TOWARDS AUTOMATIC FOOD INTAKE MONITORING USING WEARABLE SENSOR-BASED SYSTEMS

A Dissertation
Presented to
The Academic Faculty

By

Temiloluwa O. Olubanjo

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
August 2016

TOWARDS AUTOMATIC FOOD INTAKE MONITORING USING WEARABLE SENSOR-BASED SYSTEMS

Approved by:

Dr. Maysam Ghovanloo, Advisor
Associate Professor, School of ECE
Georgia Institute of Technology

Dr. Gregory D. Abowd
Professor, School of Interactive Computing
Georgia Institute of Technology

Dr. Elliot Moore II, Co-advisor
Associate Professor, School of ECE
Georgia Institute of Technology

Dr. Thad Starner
Professor, School of Interactive Computing
Georgia Institute of Technology

Dr. Omer Inan
Assistant Professor, School of ECE
Georgia Institute of Technology

Dr. Fatih Sarioglu
Assistant Professor, School of ECE
Georgia Institute of Technology

Date Approved: July 25, 2016

To God, the One I live for, who is my Savior, Provider, and Guide.

ACKNOWLEDGMENTS

This document will not be complete without me acknowledging people who have played a vital role in my life and journey thus far. First and foremost, I thank my husband, brothers and parents for their steady love, many sacrifices and unwavering encouragement. My Daddy deserves special recognition for being my loudest fan. My brother, Olutide Olubanjo deserves special recognition for our many conversations that keep me going and calling me "Ph.D." since the day I was accepted into the program. My husband, San'Quan Prioleau deserves special recognition as my best friend, my confidant and my prayer partner. To my Ph.D. Supporters Inc., thank you for being my reviewer on every scholarship/fellowship application, conference/journal paper, and general write-up I have done while in graduate school. To my friends and Ph.D. accountability team, thank you for setting an example for me from the time I arrived at Georgia Tech, naive and overwhelmed by the many challenges the Ph.D. road had in store.

I thank my advisors, Dr. Maysam Ghovanloo and Dr. Elliot Moore, for taking me under their wings, teaching me and enriching my technical skills and knowledge. My advisors have invested greatly in my growth and success. I also thank my reading committee members consisting of Dr. Omer Inan, Dr. Gregory Abowd, Dr. Thad Starner and Dr. Fatih Sarioglu. They each took time to provide me with feedback as well as research and career advice. A special thank you goes to old and new members of the GT-Bionics laboratory and Dr. Moore's team who provided me with ongoing support. In addition, I acknowledge all of my research collaborators some of whom are Antoine Liutkus and Kareem Bedri.

A big and sincere thank you goes to my funding agencies, National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP), ARCS foundation, Google Anita Borg and Google UNCF scholarships as well as the Sam Chih Foundation for recognizing my research work. These agencies provided me with the liberty to focus on my Ph.D. research work without having to undertake added financial burdens.

Finally, I would like to thank several Georgia Tech faculty and staff who supported and encouraged me at different phases of my program. Through the SURE program, started by Dr. Gary May, I found my research interest and decided to pursue a Ph.D. Thank you to members of the ECE graduate office, Jacqueline Trappier, Tasha Torrence Daniela Staiculescu for always being willing to assist me with whatever questions I had regarding the Ph.D process. Thank you to Dr. Leyla Conrad for always welcoming me to her office, encouraging and supporting me, and playing a key role in me pursing an academic career starting with a post-doctoral fellowship at Rice University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The growing crisis of obesity, diabetes, eating disorders and other chronic conditions which are directly or indirectly related to unhealthy dietary habits is a primary motivation for this research work. Food intake monitoring using wearable sensor-based systems is an alternative to manual self-report methods. The goal is to quantitatively track aspects related to eating, drinking and/or any form of energy consumption in an effort to encourage healthier behaviors. A wearable system aimed at ubiquitously monitoring eating activity in daily living should be energy efficient, unobtrusive, capable of robust functionality in various recording environments and capable of estimating relevant dietary parameters in the midst of other daily activities.

In this thesis, we focus on a detailed evaluation of research work in the field to outline pros and cons of different sensing modalities and on-body sensing locations. For the various sensing modalities implemented and evaluated in literature towards automatic dietary monitoring, we identified and reported the most relevant signal processing and machine learning methods including best features for acoustic-, image-, and motion-based methods. We delve more into acoustic sensing of food intake activities to develop the first real-time swallowing detection algorithm based on tracheal acoustics and an efficient tracheal activity recognition algorithm using a sub-optimal sampling rate for energy efficiency purposes. Next, we focus on the research gap of robust functionality of acoustic sensing methods in realistic, noise-prone, recording environments. To this aim, we develop algorithms capable of target enhancement of food intake signals from noisy recordings and food intake detection in very low signal-to-noise ratio recordings. Finally, we highlight research gaps and considerations for future work in the field. This research is towards development of an energy efficient, unobtrusive and robust wearable, sensor-based food intake monitoring system. Such a system aims to provide users with quantitative dietary feedback to support improved choices in daily living.

CHAPTER 1

INTRODUCTION

According to the National Eating Disorders Association, unhealthy dietary habits affect all ages, genders, and demographics (Figure 1), and is associated with several chronic diseases. Chronic conditions account for more than 75% of health care expenses in the United States, and are the leading cause of deaths and disabilities [22]. Preventive measures and better management can mitigate adverse effects associated with several chronic illnesses. Amongst the major conditions that affect the U.S. are obesity, eating disorders and diabetes, all of which are significantly affected by dietary behavior. Therefore, food intake monitoring is a promising research direction in the fight against obesity, eating disorders, and associated health conditions.

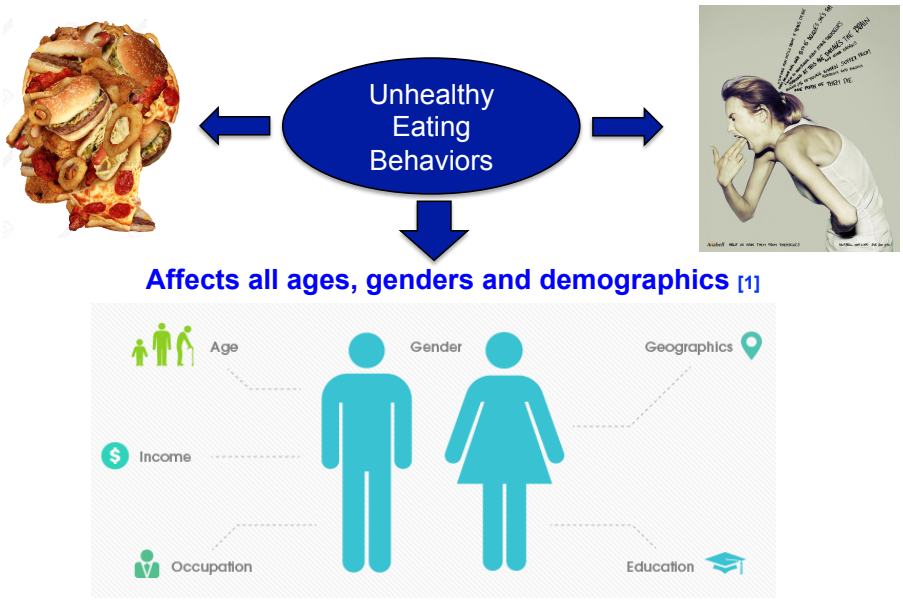


Figure 1: Problem - Unhealthy eating behaviors

Obesity alone is known to increase an individual's risk for type II diabetes, heart disease, high blood pressure, arthritis-related disability, stroke and some types of cancer [22, 23]. Obesity affects more than 1 in 3 adults and more than 1 in 6 children and adolescents from ages 6 to 19 [23]. For obesity prevention, monitoring physical activities

(or energy expenditure) in daily living is only half the battle because food intake plays a notable role in maintaining energy balance. Weight gain (or weight loss) often results from an energy imbalance; over time, when people eat and drink more calories than they burn, the energy balance tips towards weight gain and vice verse for weight loss [23].

On the other end of the spectrum from obesity is extreme weight loss caused by eating disorders such anorexia and bulimia. Eating disorders affect up to 30 million people in the U.S. [24]. Of the four modifiable health risk behaviors, regular physical activity and good nutrition from food intake can lower the risk of many of the top chronic conditions [22]. As seen from the examples in Figure 2, continuous monitoring using wearable systems has shown to be a reliable approach for activity and health tracking. Various methods have been developed for accurate and objective characterization of physical activity [25, 26], as well as estimating energy expenditure [27]. Meanwhile research efforts are in progress for monitoring dietary behavior, and there is currently no accurate and non-obtrusive means to objectively monitor food intake in daily-living conditions [11, 28].

Figure 2: Solution - Wearable systems for continuous monitoring. Commercially available examples: a) Fitness trackers, b) BioStampRC, MC10, c) Sensing food dynamics, Moticon

1.1 Major Contributions of this work

Based on the research status towards automatic food intake monitoring using wearable sensor-based systems, this work explored several avenues in this field. These are highlighted as the major contributions:

1. The first real-time swallowing detection algorithm based on tracheal acoustics [29] presented in literature. Since swallowing is a key event that always occurs during solid or liquid food intake, the ability to detection this activity in the midst of other common tracheal events is a major accomplishment. This work on real-time swallowing detection can potentially be used to trigger a more detailed, power-hungry sensor in a multi-modal wearable system such as a camera for dietary monitoring. When the frequency of swallows increases, it may be assumed that the user is eating or drinking something. This can save image storage space, reduce processing efforts on retrieve images and privacy concerns associated with taking pictures at a fixed time interval throughout an entire day.
2. An efficient tracheal activity recognition algorithm [30] for an acoustic-based dietary monitoring system. The developed algorithm focused on detecting and classifying five common and easily replicable tracheal activities namely chewing, swallowing, clearing the throat, coughing and speech. It is important to note that a sub-optimal sampling rate was used in addition to predetermined relevant acoustic features and simplistic classifiers to achieve minimal-power consumption for implementation in a wearable system.
3. Noise-handling algorithms for acoustic-based dietary monitoring systems capable of target enhancement from noisy signal [31] and food intake event detection from very low signal-to-noise ratio signal [32]. Acoustic sensors are a very common sensing modality used by researchers for food intake monitoring, meanwhile the issue of collecting and processing data from realistic noise-prone conditions was yet to be explored in previous literature. These studies are amongst the first to approach chew

event detection for food intake monitoring in recordings with up to -20 dB signal-to-noise ratio.

4. A comprehensive review of state-of-the-art research of unobtrusive sensing and wearable systems for automatic dietary monitoring. This review includes summaries of the following:

- Useful measurable dietary parameters
- Pros and cons for various wearable sensing locations
- Sensing methods including sensor-types and design criteria
- Signal processing pipeline for various sensor types as well as relevant feature types and classification techniques
- Performance benchmark of state-of-the-art dietary monitoring systems based on the study objective, sensor-type, data source, cross validation method and overall results
- Research gaps and considerations for future work towards robust automatic dietary monitoring systems

1.2 Thesis Organization

The rest of this thesis is organized as follows: chapter 2 discusses background information on food intake monitoring methods. This includes an extensive review of: 1) sensor-based dietary monitoring methods (sensor types, sensing locations and sensor combinations in single- and multi-modal wearable systems), 2) recognition methods (signal analysis and feature extraction for various sensor-types used), 3) bench-mark of state-of-the-art performance of automatic dietary monitoring systems. In chapter 3 and 4, research work on the acoustic profile of food intake events and developed acoustic-based algorithms for tracheal activity recognition and real-time swallowing detection are presented. Chapter 5 focuses on intake detection in real-life, noisy acoustic signals. Chapter 6 contains recommendations for future research based on identified research gaps from the extensive review work. Lastly, chapter 7 presents the conclusion.

CHAPTER 2

BACKGROUND

2.1 Food Intake Monitoring Methods

According to the *Dietary Guideline for Americans*, many factors affect and influence dietary choices and overall health of a person. Some of these are individual factors (age, gender, food intake and physical activity patterns), environmental settings (school, workplace and recreational facilities), sectors of influence (government and health care system), as well as social and cultural norms that govern thoughts, beliefs and behavior [33]. From the aforementioned influencers, monitoring food intake is amongst behaviors with the strongest evidence shown to have a positive impact on weight management [33].

Figure 3 shows a summary of dietary monitoring approaches in the literature and in practice, including manual and automated methods. Automated methods can be divided into fully-automated and semi-automated, both of which can be monitored using wearable, hand-held and environmental systems. Wearable monitoring systems can be single-sensor or multi-sensor devices designed for single- or multi-location on-body use. Regardless of the method of choice, the primary goal is to quantitatively track food intake parameters in an effort to encourage a healthier dietary behavior. The rest of this chapter discusses in further detail manual and automated methods for food intake monitoring.

2.1.1 Manual Monitoring Methods

All traditional dietary monitoring methods rely on self-report information, reported by the subjects themselves for tracking [34]. Three common self-reporting methods are: 1) 24-hour recall, 2) food frequency questionnaires and 3) dietary records. The 24-hour recall method involves daily calls from a trained nutritionist or interviewer in an attempt to collect and quantify the subject's food intake each day. Its success often depends on the subject's memory, cooperation, and communication ability, as well as the interviewer's skill-level

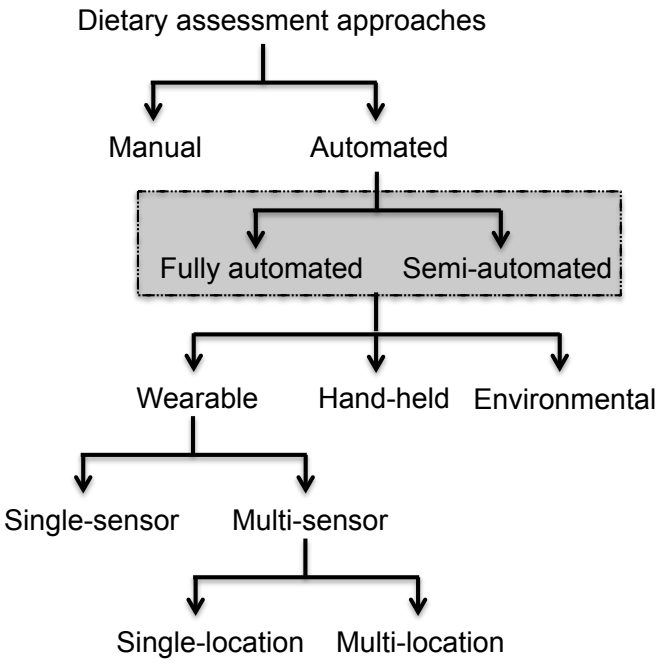


Figure 3: Breakdown of dietary monitoring approaches

[35]. Alternatively, food frequency questionnaires are often used in large cohort studies to place individuals into broad categories along a distribution of nutrient intake [34]. This method is not designed for energy intake estimation. Whereas, dietary records are detailed descriptions of the types and amounts of foods and beverages consumed, meal by meal, over a prescribed time period, usually 3 to 7 days [35]. In early literature, dietary self-reporting primarily referred to monitoring diet on paper diaries, but with more advanced technology, reporting using personal digital assistants on the internet and now smartphone applications such as *MyFitnessPal* and *MealSnap* is prominent [36].

Manual self-report methods require literate and motivated subjects [34]. The added burden these methods place on subjects could be a reason for the decline in quality of food records relative to the number of days recorded [37]. In addition, the process of actively recording food intake can cause subjects to change their eating patterns and this can lead to records containing inaccurate representation of subjects' normal food intake. Independent

of the self-reporting method used, underreporting by up to 50% is pervasive and this negatively affects efforts towards improving food intake habits and weight management [34,36].

Doubly labeled water (DLW) is another manual method often used to determine the validity of tools designed to measure energy intake [38]. When using the DLW method, subjects are given a form of "labeled" water that includes elements such as deuterium and oxygen-18, which can be measured by sampling saliva, urine or blood to estimate metabolic rate [39]. This method provides an accurate measure of a free-living subject's total energy expenditure which can be equivalent to energy intake in weight-stable individuals [34]. Due to the high cost and sophisticated technology associated with DLW, its use to date is not suitable for personal purposes as need for continuous food intake monitoring and DLW is confined to research laboratories [34].

2.1.2 Automated Monitoring Methods

Automated, sensor-based methods are being explored in research to provide a more accurate and reliable alternative for dietary monitoring. From the human body perspective, it is important to first understand specific activities related to eating that can be monitored using a sensor-based approach. Figure 4 from [1] shows eating-related activities and physiological responses to food consumption. Some of these activities are: swallowing, chewing, intake gesture, gastric activity, cardiac responses etc.

As previously stated, automated methods can be divided into fully-automated and semi-automated systems, which can further be divided into: environmental/smart-object, hand-held and wearable systems. Environmental sensing methods are fixed in predetermined locations and capture activities only in pre-determined locations [26]. Locations of interest for such systems are dining rooms, personal rooms in the home, and senior living centers as in [40]. Some environmental sensing systems in literature are video surveillance set-ups as in [41] and the meal-weighing dining table in [42].

Hand-held systems, primarily smartphones and mobile devices, have also been used towards dietary monitoring [43, 44]. In most cases, these systems allow for semi-automated

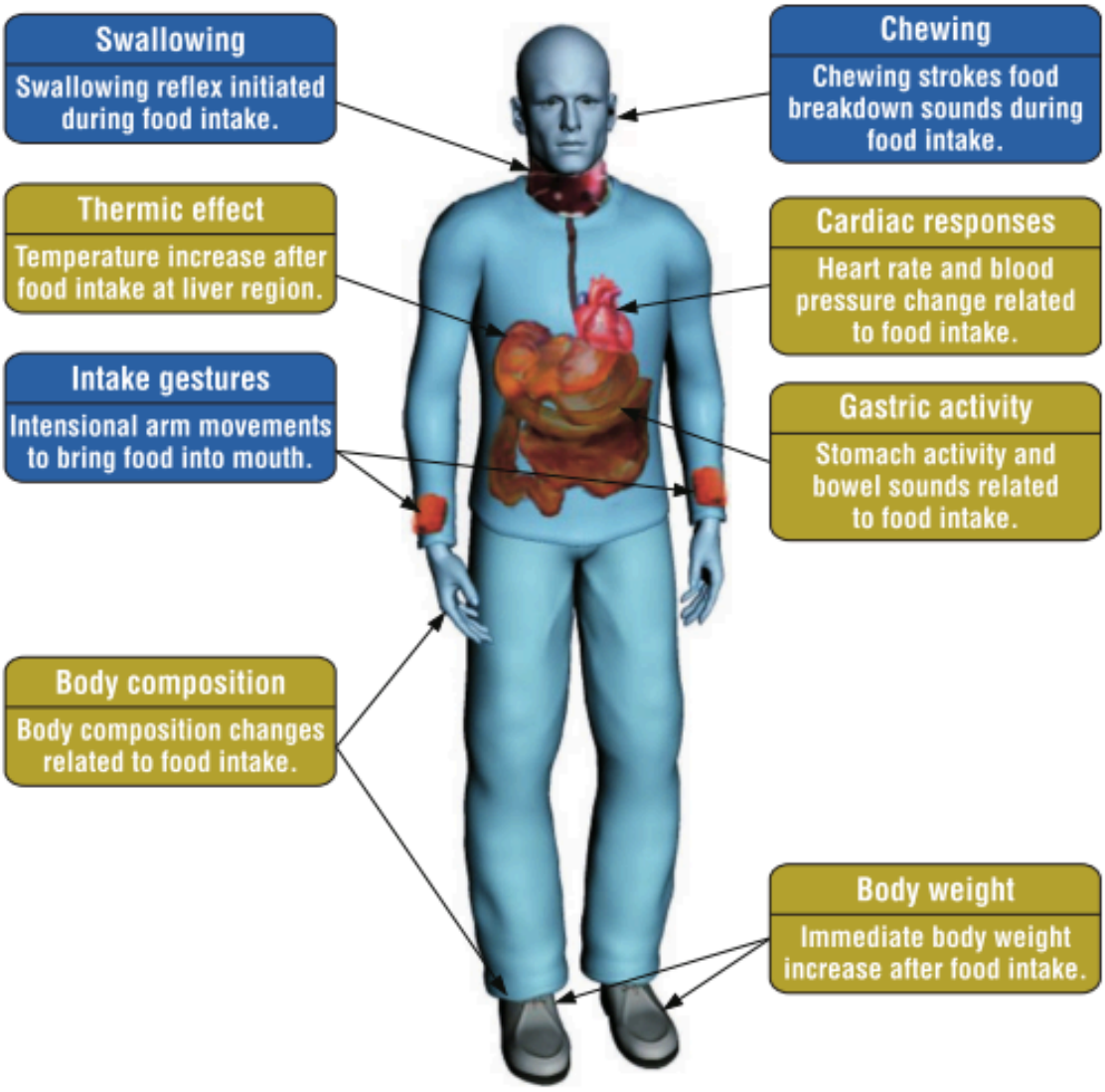


Figure 4: Eating related activities and physiological responses to food consumption [1]

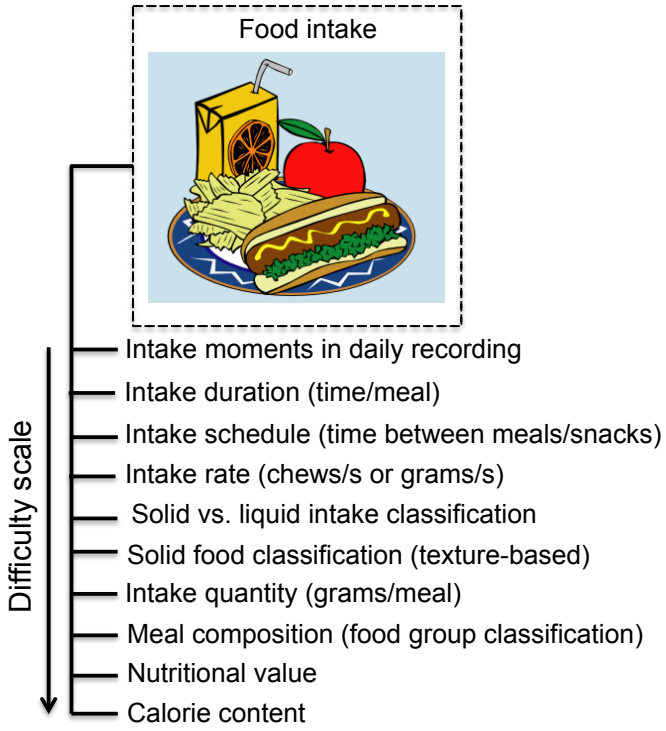


Figure 5: Measurable parameters for dietary monitoring

monitoring because they rely on the user to trigger or initiate the recording process. An advantage of smartphone-enabled monitoring is that it builds on a system that users already carry around voluntarily and it can benefit from additional sensors embedded in the device such as camera, inertial sensors, and global positioning systems (GPS). On the other hand, wearable sensors are attached to the user in one or more locations and are capable of ubiquituous sensing. For this reason, wearable systems are sometimes the preferred means for continuous dietary monitoring. To improve usuability and acceptability, wearable systems should be portable, unobtrusive, robust, privacy-preserving, flexible to support new users, energy-efficient, inexpensive and aesthetically appealing [26,45]. Common wearable methods for dietary monitoring use image recognition [11, 46], gesture recognition [15, 47] and sensors to detect chewing [21, 48, 49] and swallowing [16, 28, 50].

Figure 5 outlines several measurable parameters that can be monitored with an automated system for understanding and quantifying food intake in daily living. High-level

dietary parameters include identifying intake periods, schedule, duration and rate. Although these quantitative measures do not contain details of exactly what food item was consumed, they do provide useful information for tracking eating behavior. For example, the seemingly simple task of identifying periods of eating and intake schedule as in [21,51] can be very beneficial for seniors with dementia and alzheimer's who often forget to eat. In addition, parameters like eating rate are shown to influence meal portion size as well as expected satiety [52]. On the other hand, low-level dietary parameters include solid versus liquid intake classification, solid food classification, intake quantity, meal composition, nutritional value and calorie content. These are parameters are harder-to-obtain automatically because knowledge of the food and food type consumed is necessary. However, these parameters are important in food intake monitoring because some foods (solid and liquid) are are high in calories, but low in nutrients and can therefore leave a person overweight and malnourished [33].

It is expected that the complex problem of automatic dietary monitoring (ADM) cannot be solved using a single-sensor approach, therefore multi-modal systems should be developed in a way that each unique sensor type can contribute valuable information towards the ultimate goal of food intake monitoring even in adverse recording environments.

2.2 Review of Sensor-based Dietary Monitoring Methods

Several factors can affect the quality of a recorded signal from an ADM system's perspective including: sensor type(s) and design, on-body sensing location, recording channels and recording environment. In previous literature, ADM systems have been developed using single- and multi-sensor approaches for single- and multi-location on-body utility. Single-location systems can include one sensor as in [6, 50, 51, 53] or multiple sensors as in [2,3,19], whereas multi-location systems often include multiple sensors as in [4,18,21].

Table 1 presents the pros and cons for various on-body ADM sensing locations. Of all the locations proposed in previous work, the wrist [51, 54] is the least obtrusive primarily

Table 1: Pros and Cons: ADM Sensing Locations Proposed in Previous Literature

On-body Locations	Pros	Cons
In the mouth	i) Proximal to oral activities ii) Directly captures mouth motion	i) Least user-friendly location ii) Invasive and requires implant surgery iii) Risk of swallowing sensor unit if detached
In-ear	i) Records highest acoustic signal intensity for chewing ii) Familiar wearable location for hearing aids and earphones	i) Sensing instrument can occlude hearing
Below-ear/Behind the jaw	i) Strain sensor attachments directly monitor jaw-motion to sense chewing	ii) Strain sensor attachments require adhesion to skin; this can cause skin-irritation and sensor may lose adhesive strength during long-term recording
Neck/Throat	i) Chew and swallow acoustics are accessible from this location ii) Location can be multipurposed for monitoring other physiological events such as apnea, chronic coughing etc.	i) Often requires close sensor contact with user's neck. This can lead to a tight-fitting system around the user's neck
Wrist	i) Least obtrusive location ii) Familiar wearable location for watches and physical activity monitoring systems such as FitBit, JawBone etc.	i) Does not capture body emitted food intake sounds

because it is a familiar wearable location for watches. On the other hand, a sensor embedded in the mouth, particularly inside-the-teeth as in [12], is invasive and therefore the least user-friendly location. On-body locations such as in-the-ear and on-the-neck are familiar wearable locations for headphones/hearing aid systems and necklaces, respectively. Potential drawbacks of an in-ear device for continuous monitoring is that it can occlude hearing while a neckwear system may be tight and uncomfortable for the user. Below-the-ear/behind-the-jaw as a sensing location for continuous monitoring is not a familiar wearable location and may require the sensor to be adhered to the skin which is not comfortable or sustainable for daily use.

The most commonly monitored dietary events/items are chews, swallows, meal-images, jaw-motion, and hand-to-mouth gestures. These events have been recorded in literature through acoustic-based, image-based, motion-based and multi-modal sensing methods. The rest of the chapter presents further details on each sensing approach.

2.2.1 Acoustic-based Methods

Acoustic processing has proven valuable for other health-focused applications such as monitoring stress [55], apnea [56, 57] and cough detection [58]. This established success motivates acoustic-based methods for dietary monitoring. Acoustic recordings are contextually rich and therefore useful to gain insight on when food is being consumed as well as the type (primarily texture) of consumed food as in [2, 3, 6]. However, a primary drawback of acoustic sensing for ADM is interference of environmental and background noise. Figure 6 shows some prototypes of acoustic-based ADM systems developed by different research groups and Table 2 shows a summary of sensing locations, microphone types and number of channels used in previous literature.

2.2.1.1 Sensing Locations

Acoustic-based ADM systems have been positioned primarily in the ear [3, 53], on the neck [29, 50] and on the wrist [54]. The pros and cons of each of these sensing location are

Table 2: Summary of Acoustic-based ADM Systems in Previous Literature

Ref.	Food Intake Event of Interest	Sensing Location	Microphone Type	Acoustic Channels	Microphone Model
Nishimura et al., 2008 [59]	Chews	Ear	unknown	1	unknown
Amft, 2010 [53]	Chews	Ear	Electret, omnidirectional	1	Knowles FG23329
Shuzo et al., 2010 [2]	Chews	Ear	i) Bone conduction ii) Condenser	2	i) Vibraudio EM20, Temco ii) WM-E13U, Panasonic
Päßler et al., 2012 [3]	Chews	Ear	Electret, omnidirectional	2	Knowles FG23329-CO5
Liu et al., 2012 [19]	Chews, swallows	Ear	unknown	1	Sony ECM TL3
Sazonov et al., 2008 [4]	Swallows	Ear, throat	unknown	3	iASUS NT3, iXradio XEM98D
Yatani et al., 2012 [7]	Chews, swallows	Throat	Condenser, unidirectional	1	Custom-made
Rahman et al., 2014 [5]	Chews, swallows	Throat	Piezoelectric, unidirectional	1	Custom-made
Olubanjo et al., 2014 [30]	Swallows	Throat	unknown	1	iASUS NT3
Kandori et al., 2012 [20]	Swallows	Throat	Piezoelectric	1	unknown
Walker et al., 2014 [50]	Swallows	Throat	unknown	1	iASUS NT3
Olubanjo et al., 2014 [29]	Swallows	Throat	unknown	1	iASUS NT3
Bi et al., 2015 [6]	Chews, swallows	Throat	unknown	1	unknown
Thomaz et al., 2015 [54]	Ambient sounds	Wrist	Smartphone microphone	1	unknown

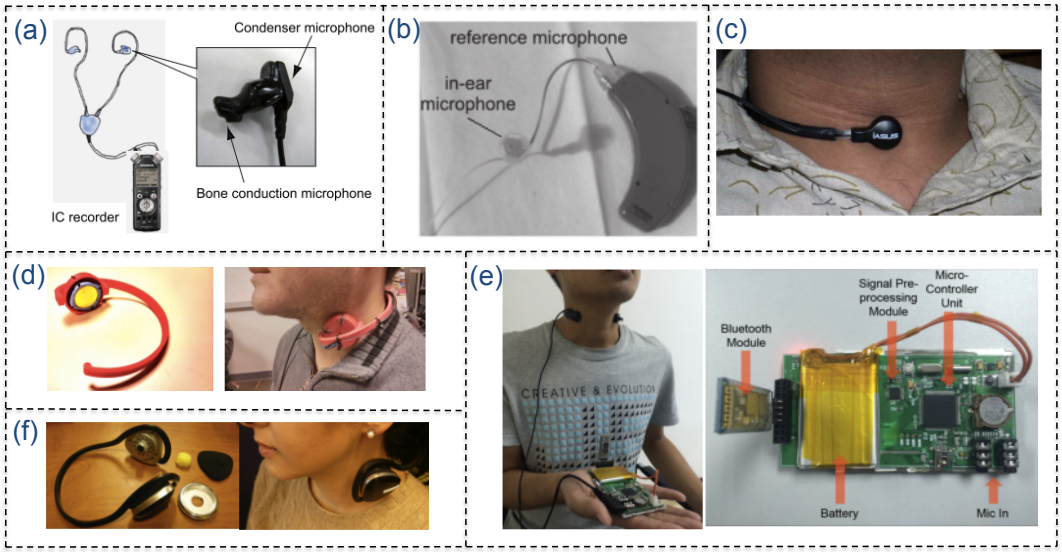


Figure 6: Examples of acoustic-based dietary monitoring systems. Sensing from the ear are (a) - [2] and (b) - [3]. Sensing from the throat/neck region are (c) - [4], (d) - [5], (e) - [6] and (f) - [7].

highlighted in Table 1. Amft et al. [60] compared the signal intensity of chewing sounds recorded from six microphone positions: inner ear, 2 cm in front of mouth, at cheek, 5 cm in front of ear canal opening, collar/neck and behind the outer ear. They found that the highest signal intensity for chewing sounds was accessible from the inner ear followed by in front of the mouth and then the collar/neck position. Of these positions, only the inner ear and collar/neck allow for wearable sensing of dietary behavior. Likewise, Rahman et al. [5] compared the recorded signal power of five activities (eating, drinking, breathing, coughing and speaking) from three microphone positions (jaw, skull and neck). They found the neck to be the best of the three locations for all activities tested except chewing. The highest recorded signal during drinking, which is a form of dietary intake, was obtained from the neck region. Meanwhile, the highest recorded signal for chewing was recorded from the 'skull', which in their paper refers to a position behind the earlobe, right around the mastoid bone. The findings from [5, 60] both support that the highest signal power for chewing (which is representative of solid food intake) can be recorded by capturing sounds propagated through bone conduction to the ear and around the mastoid bone. Whereas, the

highest signal power for drinking (which is representative of liquid intake) can be recorded by capturing swallowing sounds from the throat region. It is important to note that the maximum power for vocal and other non-vocal (breathing and coughing) activities was also recorded from the throat region [5]. Therefore, this sensing location may be appropriate for a wearable multi-modal health monitoring system useful for dietary, pulmonary and maybe even cardiac monitoring.

2.2.1.2 Microphone types

Common microphone types used in acoustic-based ADM systems are: condenser and piezoelectric microphones. In [4], Sazonov et al. compared the sound quality of four commercially available microphones: 1) piezoelectric bone-conduction (EM-L from Temco Inc), 2) piezoelectric noise-canceling (N4530 from Challenge Electronics), 3) a modified throat microphone (XTM70V from iXradio), and 4) throat microphone (iASUS NT from iASUS Concepts Ltd.). The microphone tests included subjective listening, objective visualization, and signal-to-noise ratio computation of recordings for several consecutive swallows. It was not reported whether the test swallows were spontaneous, liquid-intake, or solid-intake swallows. The authors concluded that the throat microphones showed less sensitivity to ambient noise.

In [5], Rahman et al. compared seven microphone design configurations including brass and film piezoelectric sensor-based designs with latex and silicone diaphragm materials, a condenser microphone with plastic diaphragm, and two off-the-shelf bone conduction microphones. Each microphone design was evaluated with respect to sensitivity in 20 Hz - 16 kHz frequency range and susceptibility to external (white, babble, traffic and conversational) noise. They found that for a contact throat microphone, the piezoelectric design was less susceptible to external noise than the condenser and bone-conduction designs. Additionally, of two piezoelectric designs, the microphone with latex diaphragm was slightly better to minimize external noise but the microphone with silicone diaphragm was

notably better for transferring in-body vibrations below 2 kHz. Therefore, the overall microphone comparison experiment results in [5] suggest a piezoelectric-based microphone with silicone diaphragm to be optimum for recording tracheally accessible, non-speech body sounds (including dietary sounds). Results from [4, 5] do not show consistent results that support use of a specific microphone type for optimum acoustic signal quality in ADM systems.

2.2.1.3 Recording Channels

Most acoustic ADM systems in the literature use single-channel recording [5–7, 50, 53, 59]. A few papers suggest and implement multi-channel recording for the purpose of noise reduction [2, 3, 61]. Noise handling is an important step for a robust acoustic-based ADM system capable of good functionality in various recording environments. Effective hardware design can be used to minimize external noise interference prior to the signal reaching the microphone as seen in [5]. Yet, noise reduction or target enhancement remains a necessary pre-processing step for the acquired signal which is still likely to contain some background noise. Liutkus et al. [31] proposed a single-channel target enhancement approach that learns spectral patterns of food intake acoustics from a clean signal and uses learned patterns to isolate the signal of interest from a noisy/mixed signal.

Unlike the single-channel approach in [31] which may afford lower power consumption, Päßler et al. presented a two microphone channel ADM system in [3]. The system includes an in-ear microphone primarily for recording sounds emitted from the skull bone (chewing sounds) and a reference microphone placed behind the ear primarily for recording environmental sounds. Using these two synchronous microphone channels, a ratio of the sum of absolute signal amplitude from the in-ear signal and the reference signal was computed in consecutive frames and compared to an adaptive threshold for food intake detection. This method was used to distinguish between sounds generated inside the user's body versus environmental sounds. However, it is not clear from this paper how effective their proposed method is for detecting food intake activity in a noisy signal. The data collection

systems implemented in [4] and [49] use multi-channel acoustic recordings and highlight the importance of using one microphone primarily for recording ambient noise but none of these papers presented a source separation method capable of food intake detection in a noise saturated signal (SNR $< 1dB$) as shown in [31].

There is often a trade-off between the number of recording channels, form factor and power consumption. More recording channels can increase the power consumption and form factor of a wearable ADM system which is generally undesirable. Meanwhile, more recording channels provide additional sources from which the signal can be analyzed.

2.2.2 Image-based Methods

Figure 7 shows prototypes of image-based dietary monitoring systems. These systems rely on visual cues to supplement traditional self-report methods or to achieve fully-automated monitoring using a single-sensor or multi-modal approach. In previous research, two primary platforms are used for image-based dietary sensing: 1) hand-held devices such as personal digital assistants (PDA), smartphones or tablets [9, 44, 62–64], 2) on-body wearable systems that include a camera [18, 19]. ADM systems on smartphone platforms can benefit from various in-built sensors and capabilities such as camera, global positioning system (GPS), inertial tracking, high-speed processing and wireless connectivity. However, there is a potential drawback of inconsistent image quality from different smartphones/mobile devices for image-based methods. Sharp et al. [65] highlight four dietary recording methods that use mobile phone platforms, namely: electronic food diary, food photograph-assisted self-administered 24-hour recall, food photograph analysis by a trained dietitian and automated food photograph analysis. Considering the scope of this paper, we focus only on systems that implement food photograph analysis by a trained dietitian e.g. Nutricam (Alive Technologies Pty Ltd) and automated food photograph analysis.

In [66], Gemming et al. categorize dietary image-based systems as either active or passive. In the active sensing case, a user is required to initiate the meal-image recording process by for example, taking a picture of the food per a specified protocol. Three common

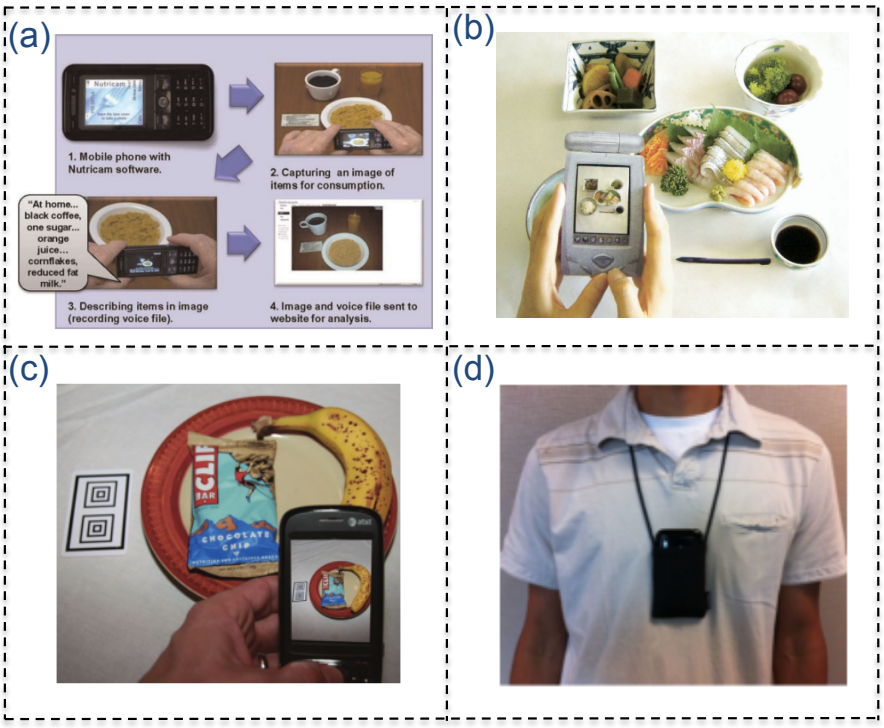


Figure 7: Example of image-based dietary monitoring systems. Sensing with mobile devices are (a) - [8], (b) - [9], (c) - [10]. Sensing with a wearable camera is (d) - [11].

requirements were observed from previous literature as part of the protocol for meal-image capture in active sensing methods. First, meal-images should be captured at a specified angle, most commonly 45 degrees as in [9, 67]. Second, for quantifying of food intake, images of food selection before eating and leftovers after eating are necesssary as in [68]. Third, a visual reference object such as a PDA stylus [9], reference ruler [67] or printed pattern [10, 69] is required to be included in meal-image pictures to facilitate estimation of parameters such as size, area and color. The system in [18] was capable of projecting a light pattern on the meal-plate for use as a dimensional referent to calculate food portion size.

In the passive sensing case, a camera can be embedded in a wearable system or positioned to capture images (or videos) from a fixed location in an environment of interest such as a dining room. In these cases, the camera is automatically activated on a fixed

time-basis such as in [11] or by detection of other activities like chewing as in [19]. Although passive sensing systems do not have the added burden of requiring users to capture meal-images, these systems have a higher probability of automatically capturing images or videos of other things/people in the scene and can therefore violate privacy. In [11], first-person images were captured every 30 s from a phone camera worn around the neck like a pendant. Due to the fixed timing for automatic image capture, such a system will require large memory capacity and high power consumption. In addition, due to possible privacy concerns of image-based passive sensing methods, the study in [11] allowed an intermediary step for users to review the entire image set and delete compromising or private images they did not want to share. In [18], identification of eating episodes from ambient sound data was used to segment meal times in continuous video recording. Then, the video dataset was automatically scanned to identify and blur-out human faces captured during recording. Such privacy measures are particularly necessary for passive image-based methods.

2.2.3 Motion-based Methods

We define motion-based ADM systems as devices that used a sensor to record and monitor a body-motion related to dietary intake. Figure 8 shows some prototypes of motion-based dietary monitoring systems. Whereas, Table 3 provides a summary of sensor types, location and events of interest from previous work on motion-based ADM systems. Sensor types used include accelerometers for sensing teeth-motion [12] and wrist-motion [51], gyroscopes for sensing wrist-motion [15], and piezoelectric sensors for monitoring jaw-motion [14, 21].

2.2.3.1 Sensing location

Motion-based ADM systems have been used on different body locations including, inside the mouth, below the ear, on the wrist and neck. Based on these locations, different body-motions are sensed to infer food intake. To the our knowledge, no work has explored the best/optimum location for motion-based ADM systems. Unlike on-body sensing methods,

Table 3: Summary of Motion-based ADM Sensing Methods in Previous Literature

Ref.	Body-motion Sensed	Food Intake Event of Interest	Sensor Type	Sensing Location	Sensor Model
Amft and Tröster, 2008 [70]	i) Throat-motion ii) Hand-motion	i) Swallows ii) Hand-to-mouth	i) Electromyo-gram ii) Accelerometer, gyroscope, compass	i) Neck ii) Arm and wrist	i) Nexus-10, MindMedia ii) MTi, XSens
Sazonov et al., 2008, 2012 [4, 14]	Jaw-motion	Chews	Piezoelectric film sensor	Below outer ear	unknown
Dong et al., 2012 [15]	Wrist-motion	Hand gesture	MEMs gyroscope	Wrist	LPR530al, STMicroelec-tronics
Li et al., 2013 [12]	Teeth-motion	Chews, Drinking	Tri-axial accelerometer	Embedded-in-teeth	unknown
Dong et al., 2014 [51]	Wrist-motion	Hand gesture	Accelerometer, gyroscope	Wrist	iPhone 4
Fontana et al., 2014 [21]	i) Jaw-motion ii) Wrist-motion	i) Chews ii) Hand-to-mouth	i) Piezoelectric sensor ii) RF transmitter	i) Below outer ear ii) Wrist	i) LDT0-028K ii) unknown
Kalantarian et al., 2014, 2015 [13, 71]	Throat-motion	Swallows	Piezoelectric sensor	Neck	LDT0-028K
Farooq et al., 2015 [72]	Jaw-motion	Sucking (breast- & bottle-feeding)	Piezoelectric film sensor	Below the ear	DT2-028K, Measurement Specialities Inc.

Figure 8: Example of motion-based dietary monitoring systems. Sensing teeth-motion is (a) - [12], throat-motion is (b) - [13], jaw-motion is (c) - [14] and wrist-motion is (d) - [15].

Li et al. [12] proposed embedding a tri-axial accelerometer in the mouth, specifically in the teeth, to take advantage of this location's close/direct proximity to oral activities. An obvious drawback of a sensor embedded inside the teeth is that it is invasive, can be affected by saliva in the mouth, and presents a risk of the user swallowing the sensor unit if it becomes detached during use. Other studies present motion-based on-body ADM systems that use a piezoelectric sensor attached below the ear (behind the mandible) to sense jaw-motion for monitoring chewing [4, 14, 21] and infant sucking [72]. An advantage of the below-ear sensing location is that it provides direct access to the lower jaw, which is involved in sucking and chewing. On the other hand, a drawback of this method is that the piezoelectric sensor in [4, 14, 21, 72] is adhered to the skin which may not be comfortable, can cause skin-irritation, and may lose adhesive strength during long-term use.

In [13, 71], Kalatanrian et al., also used a piezoelectric sensor but they propose placement against the throat to sense muscular contraction that occurs with swallow events. A potential drawback of this sensing method and location is that extraneous motion artifacts

associated with normal head and body movements can drown out the low-energy swallow signal. Also, men have more prominent hyoid and laryngeal elevation during swallowing than women [73], this may lead to a lower quality signal recorded for female users and a potentially gender-biased performance. Additionally, overweight/obese individuals have more neck adipose which may decrease the quality of recorded signal from a surface motion sensor. This in turn may lead to poorer performance for this population. Another on-body motion-based ADM approach positions accelerometers and/or gyroscopes on the wrist to sense a unique linear and rotational motion associated with biting or transferring food into the mouth with the hand [15, 51]. A benefit of the wrist as a sensing location is that it is unobtrusive because it is a familiar wearable location for watches and physical activity monitoring systems. On the other hand, for ADM the wrist does not provide access to capture equally useful and possibly more informative body generated food intake sounds.

2.2.4 Other Unobtrusive ADM Methods

Table 4 presents a summary of other non-invasive and unobtrusive dietary sensing methods. Some of these systems are not wearables such as the diet-aware dining table [42] and smart-cup [74] while others use embedded sensors in wearable systems such as the magnetic coil [20], piezoelectric respiratory belt [17], electroglottograph device [16] and proximity sensors [75, 76]. In [42], Chang et al. augmented a dining table with weighing and radio-frequency identification (RFID) sensors to monitor food movement path between tabletop containers and individuals. With the assumption that food containers are RFID tagged, the dining table can obtain nutritional information about each food. This sensing approach divides the tabletop into multiple cells/units with unique weighing sensors and assumed that each food item is correctly placed in a unique tabletop cell. A similar work that augments objects in a user's environment for dietary monitoring is presented in [74]. Lester et al. [74] focus on sensing and classifying liquid in a smart-cup using optical spectrometry and pH/conductivity probes. In their optical setup, liquid in a container is illuminated with a controlled light source and parts of the light spectrum is absorbed based on chemical

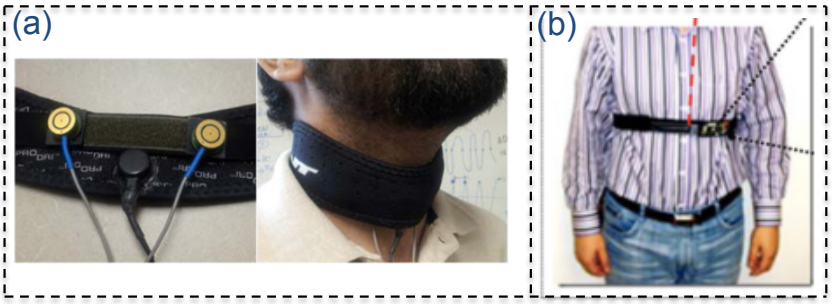


Figure 9: Example of other ADM system approaches. Sensing with an electroglottograph device is (a) - [16] and with a piezoelectric respiratory chest-belt is (b) - [17].

composition of the liquid. Additionally, a pH probe and conductivity probe is used to measure H+ ions and ability of the liquid to conduct electric current based on the salinity of the drink, respectively. Their proposed set up requires separate containers for measuring pH and conductivity to mitigate interference between emitted signals.

Unlike the smart-object approaches presented in [42, 74], Farooq et al. [16] measure laryngeal elevation for swallowing detection using a neckworn electroglottograph device. On the other hand, Dong et al. [17] base their proposed system on the observation that during swallowing there exists a short apnea which interrupts continuous breathing. The proposed system uses a piezoelectric respiratory chest belt for swallowing detection. Towards dietary monitoring, it is not known whether/how the proposed systems in [16] and [17] can differentiate between spontaneous swallows and food or liquid intake swallows. Whereas in [75] and [76] the authors use three proximity sensors (side, bottom and inner) in the ear to detect ear canal deformation during chewing. Figure 9 shows the prototypes of some of the aforementioned less-popular wearable approaches for dietary monitoring.

2.2.5 Multi-modal ADM Methods

Although single-sensor prototypes have been built and used towards dietary monitoring, another approach taken by a few research groups is to combine different sensor types into a multi-modal system. A multi-sensor approach should combine sensors in a way that it can benefit from the strength of each unique sensor included in the system. Prototypes

Table 4: Summary of Other Unobtrusive ADM systems in Previous Literature

Ref.	Event of Interest	Food Intake Event	Sensor Type	Wearable Location	Sensor Model
Chang et al., 2006 [42]	Meal weight & container RFID tag	n/a	Weighing & RFID-embedded dining table	n/a	Custom-made
Lester et al., 2010 [74]	Liquid intake	n/a	Smart-cup (optical spectrometer & pH/conductivity sensor)	n/a	Custom-made
Kandori et al., 2012 [20]	Thyroid cartilage movement	Swallow	Magnetic and acoustic sensor	Neck	Custom-made
Dong et al., 2014 [17]	Apnea	Swallow	Piezoelectric respiratory belt	Chest	unknown
Farooq et al., 2014 [16]	Laryngeal elevation	Swallow	Electroglottograph sensor	Neck	EGG-D200, Laryngograph Ltd.
Bedri et al., 2015 [75, 76]	Ear canal deformation	Chew	Proximity sensor	Outer ear	unknown

have been built in a single-unit wearable system such as [18–20] or multi-unit wearable systems such as [16]. Figure 10 shows some examples of multi-modal dietary monitoring systems from literature. From a usability and acceptability perspective, a single-unit (single-location) wearable system is preferred over a multi-unit (multi-location) wearable system.

Examples of single-unit multi-modal ADM systems are presented in [18–20]. In [19] and [18], the authors combine an image-sensor/camera with an in-ear microphone. The in-ear microphone is useful for recording and detecting chew events during eating. Chew detection is then used as a camera trigger to initiate the capture of meal images for a visual record of the exact items being consumed. It is important to consider that the meal must be in the wearable camera's field of view for this passive approach to be successful. Kandori et al. [20] combine a magnetic and acoustic sensor in a neck-worn system for swallowing detection. This system records swallowing events by monitoring the distance between two coils, one of which includes a contact piezoelectric microphone, placed on both sides of the thyroid cartilage. Unique contributions of the magnetic and acoustic sensor are not clear from the paper. Both sensors are used specially towards recognizing swallowing events.

Fontana et al. [21] present a multi-unit wearable system that combines three modalities for dietary monitoring: 1) piezoelectric sensor placed below the ear for jaw motion sensing during chewing, 2) RF-transmitter and -receiver worn on the inner wrist of the dominant arm and on a lanyard around neck respectively, for sensing hand-to-mouth gestures, 3) accelerometer in an Android smartphone for sensing ambulation. Their work used sensor fusion analysis from jaw motion and hand-to-mouth gesture sensors to detect food intake periods in a continuous 24-h recording, a major research accomplishment that only one other work by Dong et al. [51] has presented. However, an obvious drawback of the work in [21] is that is requires the user to wear three separate units on different body locations for proper functionality.

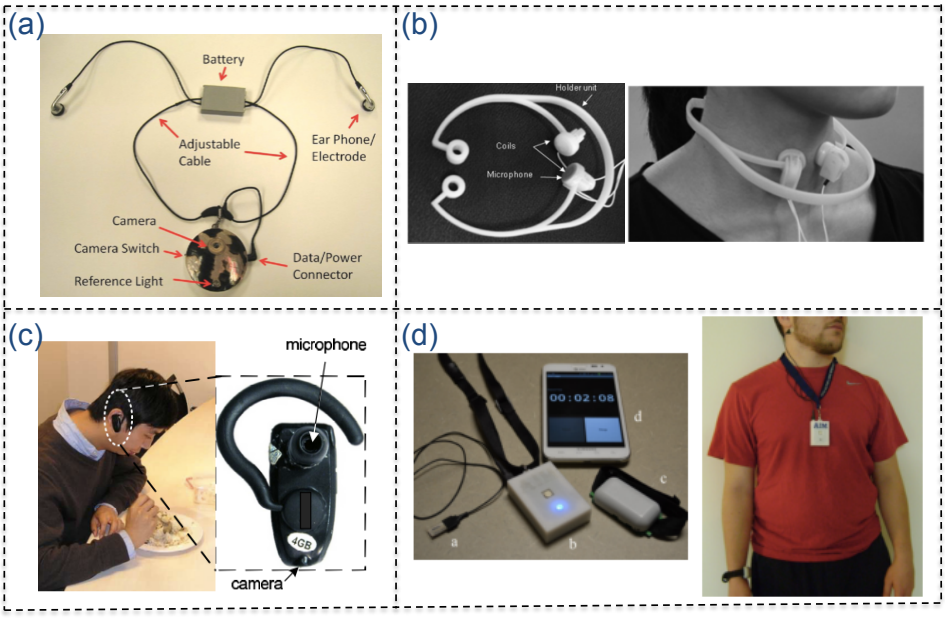


Figure 10: Example of multi-modal dietary monitoring systems. Sensing with image + acoustic sensors is (a) - [18] and (c) - [19], with a magnetic + acoustic sensors is (b) - [20], and with a piezoelectric + RF transmitter + accelerometer is (d) - [21].

2.3 Review of Automatic Dietary Monitoring Recognition Methods

This chapter includes a comprehensive literature review of signal analysis and machine learning methods for ADM systems [77]. The objective is this work is to identify the most relevant features and classifiers useful for acoustic-based, image-based and motion-based ADM systems. In addition, multi-modal signal analysis methods are presented and reviewed for identifying approaches for sensor-/data-fusion.

Overarching recognition goals of an ADM system include: 1) detect food intake activities/events in a continuous recording, 2) classify and quantify food intake activities/events, 3) extract relevant dietary parameters. All measureable parameters highlighted in Figure 5 can be categorized under one of the aforementioned three goals. The selected sensing method as described in chapter 2.2 outputs a raw signal that should be further processed and analyzed to extract relevant dietary information.

In [78], Bulling et al. provide a tutorial on human activity recognition using body-worn sensors. Although, they focus specifically on inertial sensors, the general signal processing

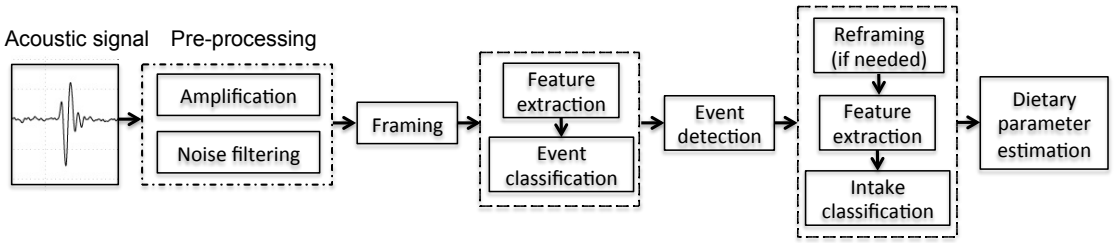


Figure 11: Acoustic processing pipeline for dietary monitoring

approach is similar for all sensing modalities. The appropriate signal processing method for an application is highly dependent on the dataset to be processed, for example, 1-D acoustic signals or 2-D image data.

2.3.1 Acoustic-based ADM Signal Analysis

An acoustic-based ADM system should be reliable in recognizing acoustic food intake events from amongst other activities in daily living such as speaking, coughing, laughing etc. Figure 11 shows a general pipeline that has been used to process acoustic signals for dietary monitoring. First, the acoustic signal is pre-processed, which can include amplification because of the relatively low energy of the signals of interest, and noise filtering to reduce extraneous noise from the background or recording environment. The pre-processing step is followed by framing, which refers to partitioning the continuous signal into smaller segments for extraction of quantitative descriptors (known as features). A first set of features can be extracted and used for event detection, which involves detecting frames with activities/events for further analysis and frames with no activities (e.g., silent frames) that can be immediately discarded. After event detection, another feature extraction step can be implemented to collect descriptive features of the food intake events of interest. These new features are used to train a classifier for intake event classification, e.g., classifying chewing and swallowing from non-food intake activities such as coughing, laughing, speaking, breathing, etc. The final step in the pipeline is estimation of dietary parameters which can be swallow count for food volume estimation or chew count for intake rate calculation.

Table 5 summarizes sampling and analysis parameters for acoustic ADM systems in

Table 5: Sampling and Analysis Parameters for Acoustic ADM Systems

Ref.	Sampling Freq. (kHz)	Frame Size (s)	Classification Window size (s)
Nishimura et al., 2008 [59]	8	0.02	n/a
Amft 2010 [53]	8	unknown	0.5
Shuzo et al., 2010 [2]	48	1	3
Paßler et al., 2012 [3]	11.025	0.023	n/a
Liu et al., 2012 [19]	44.1	0.5	3
Yatani et al., 2012 [7]	22.05	0.186	n/a
Rahman et al., 2014 [5]	8	< 0.256	1 - 5
Walker et al., 2014 [50]	44.1	unknown	unknown
Olubanjo et al., 2014 [30]	16	0.063	n/a
Thomaz et al., 2015 [54]	11.025	0.05	10
Bi et al., 2015 [6]	8	0.5	unknown

previous literature. The sampling rate must be set high enough to maintain important characteristics of the signals of interest while minimizing power consumption from the limited battery source in a wearable system. Sampling frequencies ranging from 8 - 44.1 kHz have been used in previous work. Paßler et al. [3] successfully used acoustic signals recorded at a sampling frequency of 11.025 kHz from the ear to classify 8 food types from chewing sounds, while Olubanjo et al. [30] showed that 16 kHz is a sufficient sampling rate to discriminate non-food intake events (coughing, clearing throat and speaking) from food intake events (chewing and swallowing). Rahman et al. [5] used acoustic signals sampled at 8 kHz for discriminating food intake from non-food intake activities. Meanwhile, Amft [53] and Bi et al. [6] classified various food types based on the chewing sounds also sampled at 8 kHz. These studies suggest that an acoustic sampling rate of 8 kHz could be sufficient and $>$ 8 kHz is not an effective use of power from a limited battery source for a wearable ADM system.

Chews and swallows are relatively low energy signals, therefore a pre-processing step that includes amplification can be beneficial as in [4, 6, 48]. As discussed in section 2.2.1.1

and 2.2.1.2, it is important to note that the amplitude of food intake signals recorded with a wearable system depends on the sensing location chosen and the microphone type used. Another pre-processing step necessary for realistic ADM systems is noise filtering, as in [3, 6, 59]. Although acoustic recording systems can and should be uniquely designed to minimize environmental noise interference, as in [5], dietary monitoring in a loud restaurant environment for example, would still include interfering background noise. In [59], the authors use a 4th-order Butterworth filter with a cut-off frequency of 2 Hz applied to the log energy signal for filtering. Paßler et al. [3] employed a method similar to spectral subtraction for noise-handling using concurrently recorded signals from a reference microphone and an in-ear microphone. Liutkus et al. [31] used a semi-supervised non-negative matrix factorization (NMF) to separate clean chewing sound from real-life restaurant background noise mixed at varying signal-to-noise (SNR) ratios in the range of $[-20,\ 10]\ dB$. Results in [31] show up to 20 dB improvement in separation quality in very low SNR conditions of $[-20,\ 5]\ dB$ and $\sim 60\%$ increase in chew event detection when comparing the performance on the estimated clean signal versus the raw noisy signal. Alternatively, Olubanjo et al. [32] did not focus on extracting the target (clean) signal but on detecting chew events in the noisy signal using template-matching and sliding window correlation. Results in [32] show detection performance with an F_1 score of 71.4% in very low SNR ratio signals of -10 dB compared to the 19.2% when using the maximum sound energy algorithm proposed in [49].

Another important step for ADM signal analysis is framing which refers to selection of an appropriate window size for feature extraction. The size of this feature extraction window depends on the activity of interest for recognition or classification. Unlike in speech recognition where 25 ms is the standard frame size, no standard frame size has been widely accepted for acoustic detection of food intake events. Table 5 shows that various frame lengths have been used in previous literature ranging from 16 ms - 1 s. Based on the average duration of chew, $\sim$ 0.3 s [59, 70], and average duration of swallow, $\sim$ 0.5 s [79, 80],

a frame size of > 0.3 s is not ideal because it may not describe the events of interest with small enough granularity. It is common to implement overlapping windows/frames (e.g. 50% overlap) to minimize edge effect.

Descriptive features in time, frequency, cepstral and other domains can then be extracted for each frame. Table 6 shows a summary of feature extraction methods from previous work for acoustic-based ADM systems. According to [5, 6, 19], particularly relevant features for acoustic recognition and classification in dietary monitoring systems (marked with a * in Table 6) are time domain features: peak-value, zero-crossing rate, short-time energy and energy entropy, and frequency domain features: maximum and mean power, subband power and spectral flux. After feature extraction, statistical descriptors (e.g. mean, maximum) can be used to further describe feature vectors in defined classification windows. Table 5 also shows varying classification window sizes used in previous literature range from 1 - 10 s. A feature selection step can be implemented to discover relevant and non-redundant features from the entire feature set. Paßler et al. [3] used prinical component analysis and Rahman et al. [5] used a correlation feature selection algorithm and the sequential forward feature selection algorithm. Meanwhile, Liu et al. [19] compared the performance of three feature selection algorithms namely Relief, Simba [81] and maximum relevance and minimum redundancy (mRMR) criterion [82].

Training and testing of a robust classification model is the final step for activity recognition. Table 7 shows a summary of classification methods for acoustic ADM systems. More common classifiers include nearest neighbor, support vector machine (SVM) and linear discriminant analysis. In [30], the authors compared performance of k-nearest neighor (K-NN) classifiers with Naïve Bayes and found 1-NN and 3-NN to perform better for tracheal activity classification. Whereas, in [7], the authors compared performance of Naïve Bayes, 5-NN and SVM classifiers and found SVM to be the preferred classifier also for tracheal activity classification. This is not suprising because SVM classifiers have shown

Table 6: Summary of Feature Extraction Methods for Acoustic ADM Systems

Group	Method
Time domain	Peak value*, mean, variance, standard deviation, zero-crossing rate*, energy (short-time*, entropy*, log, gap between local neighbored maximas), total variation, envelop shape statistics, skewness, kurtosis, interquartile range [2, 5–7, 19, 30, 54, 59]
Frequency domain	Maximum peak frequency, power (maximum*, mean*), ratio of band power to total power, sub-band power*, spectral centroid, spectral flux*, spectral variance, spectral skewness, spectral kurtosis, spectral slope, spectral roll-off, spectral auto-correlation, spectral autocovariance, barycentric frequency [2, 5–7, 19, 30, 50, 54]
Other	Mel-frequency cepstral cofficients, auto-regression coefficients, linear predictive coefficients, wavelet decomposition (delta coefficients), slope of detrended fluctuation analysis, approximate entropy, fractal dimension, hurst exponent, correlation dimension [5–7, 30, 50, 53, 54, 59]

Table 7: Summary of Classification Methods for Acoustic ADM Systems

Classification / Learning Methods	Threshold-based [59], Naïve Bayes [53], Nearest Neighbor [2, 30], Hidden-Markov Model [3], Neural Networks [19], Support Vector Machine [7, 50], Linear Discriminant Analysis [5, 50], Random Forest [54], Decision Tree [6]

to be robust and highly generalizable for a wide variety of datasets [83]. Dietary parameters inferred from acoustic-based ADM systems include detection of intake moments in daily recording [54], solid versus liquid intake classification [19, 50, 84], food type classification [3, 6, 53], chew count [2, 49, 59] and meal composition [85].

2.3.2 Image-based ADM Signal Analysis

Figure 12 shows a general signal analysis pipeline for image-based ADM systems. As mentioned in section 2.2.2, image acquisition is often achieved by passive or active sensing, using a wearable or hand-held device, respectively. In passive sensing cases, images are automatically captured on a fixed time basis during the day or with a wearable camera triggered by detection of other activities such as chewing. In this case, all images captured

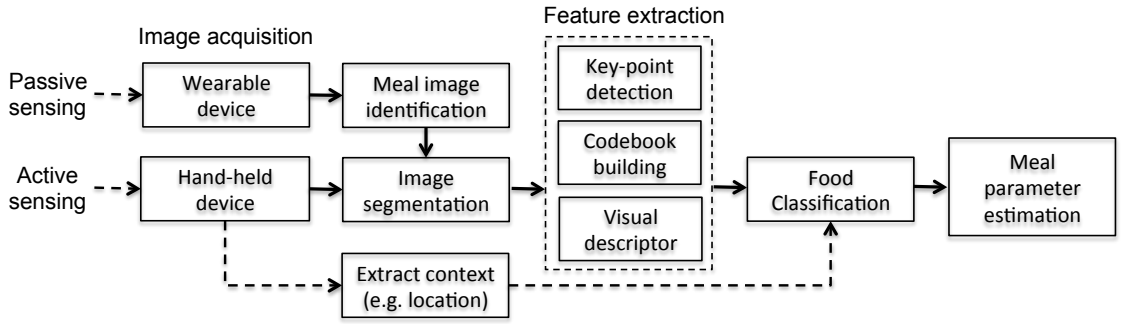


Figure 12: Image processing pipeline for dietary monitoring

are not relevant to food intake therefore a meal image identification step is needed. The goal of this step is to identify specific images that include the meal of interest. Thomaz et al. [11] implemented a coding step using Amazon's Mechanical Turk (AMT) to recognize eating moments from first-person point-of-view images. Liu et al. [19] implemented a plate search algorithm according to [86] for the meal image identification in a video sequence. Following selection of images that contain foods/meals of interest, an image segmentation step is imperative to identify specific food regions and segment food items on a plate. In [69], Zhu et al. implemented connected component analysis, active contours and normlized cuts to achieve image segmentation. The next step in the pipeline is feature extraction from image regions of interest.

Table 8 shows a summary of feature extraction methods (most relevant features are marked with a *) from previous work on image-based ADM systems. Studies in [87–89] support the relevance and effectiveness of bag-of-features (BoF) for image-based food classification. A BoF method, similar to bag-of-words used in textual information retrieval, is based on orderless collections of quantized local image descriptors independent of spatial information [90]. Primary steps necessary for BoF implementation are: key point extraction, local feature extraction, visual dictionary learning and descriptor quantization. In [90], O'Hara and Draper point out that determining the best techniques for sampling images and local image features are amongst key challenges for successfully implementing a BoF model. Anthimopoulos et al. [87] identified dense sampling as the best method for

Table 8: Summary of Feature Extraction Methods for Image-based ADM Systems

Feature Extraction Method	Descriptors
Bag-of-Features (BOF) [87–89]	SIFT (hsvSIFT*, rgSIFT, rgbSIFT, hueSIFT, opponentSIFT*, cSIFT)
Histogram [44, 87, 88, 91]	Opponent color histogram, hue histogram, gradient histogram, RGB color histogram*
Pairwise features [92]	Distance, orientation, midpoint category, between-pair-category, distance-orientation, orientation-midpoint*
Deep convolutional neural networks (DCNN) [93]	L2 normalized DCNN layer 7*
Others	Bag-of-SURF* [44], color moments [87, 91], color moment invariants [87, 91], gabor texture features [69, 89], CIELAB [69], RootHoG [93], mean and variance of RGB pixels [93]

key point extraction, while Hoashi et al. [88] identified random sampling for key point extraction. Studies [87, 88, 91] support the use of scale invariant feature transform (SIFT) as visual descriptors, more specifically hsvSIFT and opponent SIFT were shown to be highly relevant. Other highly relevant features from previous work for food image classification are pairwise features - particularly the joint pair of orientation-midpoint [92], RGB color histogram and speeded up robust features (SURF) [44] as well as the L2 normalized deep convolutional neural network layer 7 outputs [93].

After feature extraction, final steps for image-based ADM systems are food classification and meal parameter estimation. In [91], Bettadapura et al. showed the benefit of leveraging context to support automated food recognition, such as using location through geo-tags to narrow down the categories for improved food classification. All image-based ADM papers surveyed in this work used a variation of SVM classifiers (e.g. linear kernel SVM, multiple kernel SVM) in their studies. Dietary parameters inferred from image-based ADM systems include detecting intake moments during daily recording [11], food type classification [44,69,87,88,91,92], estimating food portion size [69] and some nascent

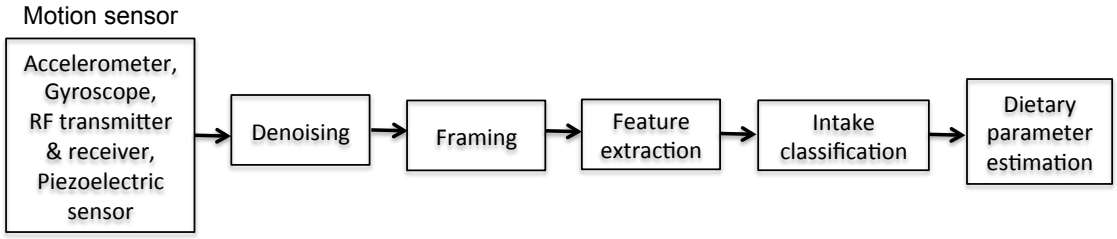


Figure 13: Motion-sensor processing pipeline for dietary monitoring

attempts at estimating calorie contents [46, 94].

2.3.3 Motion-based ADM Signal Analysis

Figure 13 shows the general pipeline for signal analysis from motion-based ADM systems. Different motion sensors have been used in previous literature including:

- Accelerometer: sensing teeth-motion [12], hand-to-mouth gesture [70], wrist-motion [51] and body-motion [21]
- Gyroscope: sensing wrist-motion and -rotation [51, 70]
- RF transmitter and receiver: sensing hand-to-mouth gesture [21]
- Piezoelectric sensor: sensing jaw-motion [14, 21, 72], throat-motion [13, 71], and chest-motion [17]

Based on the sensor type and sensing objectivity, different sampling frequencies, frame sizes, and features have been used as can be seen in Table 9. The more common sampling rate for accelerometers in previous work is 100 Hz [12, 21, 70] while a few studies have used lower sampling rates of 15 Hz [51] and 20 Hz [13]. Sampling rates for 15 Hz [51] and 100 Hz [70] have been used for gyroscopes, while sampling rates for piezoelectric sensors range from 20 - 1000 Hz [13, 14, 17, 21, 72].

Following raw data collection from the respective motion sensor, denoising is a necessary step to remove signal variations from pertubations/spikes due to short vigorous motions or noise on the power lines. In [51], accelerometer and gyroscope data for detecting hand-to-mouth gesture was smoothed using a Gaussian-weighted window while in [13], piezoelectric data for detecting swallowing from the throat region was smoothed using a

Table 9: Sampling and Analysis Parameters for Motion-based ADM Systems

Ref.	Motion Sensors	Sampling Freq. (Hz)	Frame Size (s)	Features
Amft et al., 2008 [70]	Accelerometer, gyroscope	100	0.5	Mean, variance, signal sum
Sazonov et al., 2012 [14]	Piezoelectric	100	30	Root mean square (RMS), entropy of filtered signal, base 2 log, mean, max., median, max. to RMS ratio, RMS to mean ratio, number of zero crossings, mean time between crossings, max. time between crossings, min. time between crossings, std. dev. of time between crossings, entropy of zero crossings, number of peaks, entropy of peaks, mean time between peaks, std. dev. of time between peaks, peaks to zero crossing number ratio, zero crossing to peak number ratio, entropy of spectrum, std. dev. of spectrum, peak frequency, fractal dimension
Li et al., 2013 [12]	Accelerometer	100	2.5	Mean, absolute value mean, max., min., max-min, zero crossing rate, RMS, std. dev., median, 75% percentile, inter-quartile range, inter-axis correlation, spectral entropy, energy, FFT coefficients
Dong et al., 2014 [51]	i) Accelerometer ii) Gyroscope	15	60	i) Energy peaks, manipulation (rotational vs. linear motion ratio), linear acceleration, wrist-roll motion, regularity of wrist roll motion
Dong et al., 2014 [17]	Piezoelectric	30	-	Spectral power at 3 Hz frequency bands

Table 10: Sampling and Analysis Parameters for Motion-based ADM Systems Continued

Ref.	Motion Sensors	Sampling Freq. (Hz)	Frame Size (s)	Features
Fontana et al., 2014 [21]	i) RF transmitter & receiver ii) Accelerometer iii) Piezoelectric	i) 10 ii) 100 iii) 1000	30	i) Hand-to-mouth (HtM) gestures (number in fixed time, duration, mean absolute value, std. dev., max. value) wavelength, ratios of aforementioned features ii) Mean absolute value, std. dev, median, number of zero-crossings, mean time between zero crossings, entropy iii) Mean absolute value, RMS, max., median, entropy, number of zero crossings, mean time between zero crossings, number of peaks, average range, mean time between peaks, wavelength, sub-band energy, fractal dimension, peak frequency in sub-bands
Farooq et al., 2015 [72]	Piezoelectric	1000	10	Number of zero crossings > threshold
Kalantarian et al., 2015 [13]	i) Piezoelectric ii) Accelerometer	i) 20 ii) 20	0.45	ii) Harmonic mean, geometric mean, standard deviation, kurtosis, skewness, mean-absolute deviation

sliding-window average of the original data. In [14], Sazonov et al. filtered the piezoelectric signal for detecting jaw-motion from chewing using a bandpass filter with cutoff frequencies of 1.25 Hz and 2.5 Hz. This frequency band was set based on earlier studies that determined chewing frequency to be in the range of 0.7 - 2 Hz [95]. On the other hand, Farooq et al. [72] used a bi-orthogonal wavelet transform with 4 vanishing moments to denoise piezoelectric data used for detecting jaw-motion from sucking actions of babies feeding. After the denoising step, a similar framing step is necessary as described in section 2.3.1. The appropriate frame size highly depends on the length of activity of interest. Table 9 highlights different frame sizes that have been used for motion-based ADM systems ranging from 0.45 - 60 s. In [13], the authors were interested in detecting swallow events which have an average duration of $\sim$ 0.5 s [79, 80] and they used a sliding window length of 0.45 s with maximum overlap. Whereas in [21], the authors were interested in detecting eating moments during a 24-h period and used a frame size of 30 s.

A wide range of features have also been extracted from the different sensors used for motion-based ADM systems. Details on feature extraction methods used is summarized in Table 9. Statistical features are most common such as mean, maximum, minimum, standard deviation, mean absolute value, root mean square, zero crossing rate, entropy [12–14, 21]. Feature selection is an optional yet recommended next step to minimize redundant features; forward selection procedure [96] was used in [14] while principal component analysis (PCA) [97] was used in [12].

The final step before dietary parameter estimation is intake classification. Similar classifiers used in acoustic-based ADM systems (Table 7) are applicable and have been used with motion-based ADM systems including SVM, decision tree, Naïve Bayes, and artificial neural networks. In [12], Li et al. compared the performance of C4.5 decision tree, multivariate logistic regression and SVM for 4-class activity classification (coughing, drinking, chewing, and speaking) using an accelerometer embedded in the teeth and found SVM to produce the best results. Amft et al. [70] classified 4 intake gestures (eating with fork and

knife, drinking from a glass, eating with a spoon and eating with one hand) using inertial sensors, accelerometer and gyroscope, positioned on the upper and lower arm. They found the sensors on the lower arm to be more useful and informative for intake gesture classification. Other classification problems undertaken in motion-based ADM literature include chewing versus non-chewing using a piezoelectric sensor attached to the jaw area directly underneath the earlobe [14], infant sucking count and sucking rate also using a piezoelectric sensor placed on the jaw [72] and swallowing detection using a piezoelectric belt worn around the chest [17]. Example of dietary parameters inferred from motion-based ADM systems include detecting intake moments during daily recordings [21, 51], solid versus liquid intake classification [13] and attempts at calorie estimation from bite counts [15].

2.3.4 Multi-modal Signal Analysis

As mentioned in section 2.2.5, multimodal ADM systems aim to benefit from advantages of various sensor types in a combined, possibly more robust system. Boström et al. [98] provide a comprehensive review of previous definitions of information fusion including data and sensor fusion. Whereas, Zheng et al. [45] propose a fitting definition (for this paper) of data fusion as "efficient methods for automatically or semi-automatically translating the information from multiple sources into a structured representation so that human or automated decision can be made accurately." Multisensor data fusion certainly has unique benefits and challenges. Potential advantages include improved detection, confidence, reliability, as well as extended spatial and temporal coverage [99]. Meanwhile obvious challenges, especially for a wearable ubiquitous system, include how to optimally combine hetergenous data streams and minimize power consumption. Khaleghi et al. [99] highlight other general issues related to multisensor data fusion some of which are handling conflicting data, data correlation and data alignment.

Fusion approaches can be categorized into 3 groups namely, 1) statistical approach, 2) probablistic approach, and 3) artificial intelligence [45]. Most multi-modal ADM systems in previous literature utilize the statistical approach which refers to using weighed

combinations. Fontana et al. [21] implemented an equal weighing, two-step sensor fusion approach of jaw-motion, hand-gesture and accelerometer signals. Their first fusion step created a new signal from the product of absolute values for jaw-motion and hand-gesture signals in non-overlapping 30 s window frames. Whereas their second fusion step created a new signal from the average of x-, y- and z-axis from the accelerometer signal. A new vector, created by grouping results from fusion step 1 and 2, was then used to discriminate food intake and non-food intake windows.

2.4 Benchmarking State-of-the-Art ADM systems

To enable a comprehensive summary of state-of-the-art ADM systems, results in literature were categorized into event detection and classification. The classification summary table includes papers with > 2 (binary) group discrimation such as relevant-activity classification (e.g. breathing, speaking, chewing, swallowing, coughing), food type classification (e.g. bread, cheese, vegetables, meat, pasta), texture-based food classification (e.g. dry-crispy, wet-crispy, crunchy, soft) etc.

Evaluation aspects for comparison of event detection and classification performance for ADM systems from previous work are:

- Objective (event detection or classification goal)
- Sensor(s) type (acoustic, image, motion, multimodal)
- Number of subjects used for performance evaluation
- Data source (e.g. in-laboratory experiment, real-world, online dataset)
- Cross-validation method (hold-out, *k*-fold, LOSO, LOPO)
- Overall results (accuracy, F_1 score, TPR, FPR)

2.4.1 ADM Event Detection

Event detection tasks commonly approach the high-level problem of identifying eating moments (chews, swallows and/or hand-to-mouth gesture) in a continuous recording. This can include meal consumption or sporadic snacking events. Systems capable of robust eating

detection can supplement standard self-report methods that rely on the user's memory for food tracking or to monitor eating regularity of patients/older adults. Table 11 presents a summary of event detection (binary classification) performance for ADM systems. Eating detection performance in literature ranges from $\sim 80\% - 95\%$ for controlled in-laboratory studies. Whereas, detection performance ranges from $\sim 28.7\% - 90\%$ for less controlled, in-the-wild (real-world) studies. As expected, subject independent performance is often significantly less than subject dependent performance.

A few relevant papers on event detection and binary classification for ADM systems are reviewed in further detail below:

- *Paßler and Fischer [49]:* In this work, the authors develop and test 8 algorithms for automated chew event detection in a continuous chewing sounds. Their 17-h dataset, collected from 51 subjects ranging in age from 15 to 77 years (mean: 34.8 years), includes a total of 68,094 chew events. It was recorded using a custom-built single-unit wearable acoustic system with 2 recording channels (in-ear and reference microphone). Of the 8 algorithms presented, the proposed "sound recognition" algorithm performed best on the food intake data with recall of 82% and precision of 87% (F_1 score = 84.4%). The proposed "maximum energy ratio" algorithm generated the smallest number of insertions on a mixed dataset that includes environmental noise. Potential drawbacks in this work are: i) data was collected in a controlled lab study, ii) dataset did not include any extraneous, non-chewing events that will otherwise be present in daily living recordings, iii) all algorithm presented were based on impirically defined thresholds.

- *Bi et al. [6]:* In this work, the authors present *AutoDietary,* a neckworn acoustic-based system to monitor and recognize food intake in daily living. The *AutoDietary* prototype consists of a high-fidelity throat microphone for data acquisition, an embedded hardware board for power supply, data pre-processing and transmission, a smartphone application for food type recognition, data management and visualization. Mel Frequency Cepstral Coefficients (MFCC) was used with Hidden Markov Model (HMM) for event detection in

Table 11: Summary of Event Detection Performance for ADM Systems

Ref.	Objective	Sensor(s)	# of Subj.	Data Source	Cross Validation	Result (Acc, F_1, TPR, FPR)
Paßler [3]	Eating detection	Acoustic	40	Lab	LOPO	Acc: 83.3%
Sazonov [14]	Chewing detection	Piezoelectric	20	Lab	20-fold	Acc: 80.9%
Dong [15]	Bite detection	Gyroscope	47	Lab	unknown	F_1: 83.4%
Dong [51]	Eating detection	Accelerometer, gyroscope	43	Real-world	LOOCV	Acc: 81%
Dong [17]	Swallowing detection	Piezoelectric	7	Lab	10-fold	TPR: 96.6%, FPR: 0.8%
Olubanjo [29]	Swallowing detection	Acoustic	4	Lab	Hold-out	F_1: 73.2%
Paßler [49]	Chewing detection	Acoustic	51	Lab	unknown	F_1: 84.4%
Fontana [21]	Eating detection	RF transmitter & receiver, accelerometer piezoelectric	12	Real-world	LOOCV	Acc: 89.8%
Thomaz [54]	Eating detection	Acoustic	21	Real-world	10-fold, LOPO	F_1: 79.8% (SD), 28.7% (SI)
Bi [6]	i) Eating detection ii) Solid vs. liquid	Acoustic	12	Lab	4-fold	Acc: i) 86.6% ii) 98.7%
Kalantarian [13]	Swallowing detection	Piezoelectric	30	Lab	LOOCV	TPR: 83.7%
Bedri [76]	Eating detection	Proximity	20	Lab	LOPO	Acc: 92.9% (SI)
Bedri [76]	Eating detection	Proximity	6	Real-world	LOPO	F_1: 76.2% (SI)

a continuous recording. Next, 3 categories of features (time-domain, frequency-domain, and non-linear) were extracted and used with a decision tree classifier for classification of 7 food types. Their dataset, recorded from 12 subjects ranging in age from 13 - 44 years (mean: 28.8 years), contained 4047 events (including 54 bites, 3433 chews and 560 swallows). Their proposed event detection algorithm performed with an accuracy of 86.6%, food type recognition algorithm performed with an accuracy of 87.1% and the solid/liquid classification accuracy was 98.7%. Potential drawbacks in this work are: i) data was collected in a controlled lab study, ii) dataset did not include any extraneous, non-eating activities that will otherwise be present in normal free-living recordings such as head movement, speaking, coughing etc.

- *Bedri et al. [76]:* An Outer Ear Interface (OEI) that contains a 3D gyroscope and 3 proximity sensors in an off-the-shelf earpiece was presented. The objective is to monitor jaw movement during eating by measuring ear canal deformation. Their dataset contains: 1) in-laboratory recording of 20 - 25 mins each from 20 subjects ranging in age from 18 to 41 years (mean: 24 years) as they read aloud, silently browsed the internet, ate and drank, 2) in-the-wild recording of 6-hours each from 6 subjects as they conducted daily activities of choice. Five features were used namely: 1st PCA component from proximity sensor signals, energy of proximity sensor signals and raw gyroscrope data (x-, y- and z-axis). Using hidden markov models (HMMs), eating and non-eating/null classes were trained. A subject-independent F_1 score of 92.9% was obtained on the in-laboratory dataset while a subject-dependent F_1 score of 76.2% was obtained on the in-the-wild dataset.

- *Dong et al. [15]:* The authors present a wrist-worn, watch-like sensing unit embedded with an accelerometer and gyroscope for detecting eating periods throughout the day. Their proposed algorithm starts with a preprocessing step to smooth sensor data, followed by a segmentation step to determine periods of large wrist-motion, feature extraction over interpeak segmented periods, and classification using Naïve Bayes to identify eating periods. The proposed detection algorithm in based on the observation that before and after an

eating activity, there tends to be larger wrist motion energy. Their dataset was recorded for 8.5 - 12 hours each using an iPhone 4 attached to the forearm of 43 subjects ranging in age from 18 - 50 years. The total dataset of 449 hours includes 22.4 hours of eating over 116 meals/snacks. An eating detection accuracy of 81% was obtained at 1 s resolution.

2.4.2 ADM Activity Classification

Table 12 and 13 show a summary of ADM systems that have a goal of > 2-class classification. This includes tasks such as intake gesture, food image, and food texture classification using inertial, acoustic and image sensors.

A few important papers are reviewed in further details in the below text:

- *Hoashi et al. [88]*: The authors implement an automatic food image recognition algorithm which was tested on classification of 85 food images from the web. This paper is an expansion on their previous work [89] in which they obtained classification performance of 61.34% for 50-class food image recognition. Using various image features such as bag-of-features, color histogram, gabor texture features and gradient histogram in combination with a multiple kernel learning SVM classifier, they achieved 62.52% for 85-class food image classification. They found BoF to be the most important feature for food image classification and Difference of Gaussian (DoG) point-sampling method with a codebook size of 1000 to be most effective for building the BoF vector. Furthermore, they tested the proposed algorithm on first-person images obtained with a cell-phone camera using the same 85-classes previously defined and achieved a classification rate of 45.35%. A potential drawback of this work is high variance (17% - 95%) in food image classification rates.

- *Zhu et al.* [69]: In this paper, automatic food type classification and portion estimation was developed and implemented for food images recorded from a mobile device. Meal images were segmented to determine particular food regions in the entire image using connect component analysis, active contours and normalized cuts methods. Next, food types were

Table 12: Summary of Classification Performance for ADM Systems

Ref.	Objective	Sensor(s)	# of Subj	Data Source	Cross Validation	Result (Acc, F_1, TPR, FPR)
Amft [70]	4-class gesture classification (eating with fork & knife, drinking from glass, eating with spoon, eating with one hand)	Accelerometer, gyroscope, compass	4	Lab	n/a	F_1: 71.1% (SD)
Joutou [89]	50-class food image classification	Image	n/a	i) Online ii) First-person	5-fold	TPR: i) 61.3% ii) 37.35
Yang [92]	7-class food image classification	Image	n/a	Online [100]	3-fold	Acc: 78%
Hoashi [88]	85-class food image classfication	Image	n/a	i) Online ii) First-person	5-fold	TPR: i) 62.5% ii) 45.3%
Shuzo [2]	4-class classification (eating hard food, soft, food, drinking water, speaking)	Acoustic	5	Lab	LOPO	Acc: 80% (SD), 70% (SI)
Zhu [69]	19-class food image classification	Image	n/a	First-person	10-fold	95.8%
Paßler [3]	8-class food classification (drink, pudding, choclate, walnut, peanut, carror, apple, potato chips)	Acoustic	51	Lab	LOPO	Acc: 79% (SD), 66% (SI)
Liu [19]	4-class classification (eating, drinking speaking, others)	Acoustic	6	i) Lab ii) Real world	Hold-out	TPR: i) 80.4% ii) 71.6%
Yatani [7]	12-class classification (eating, drinking, speaking, clearing throat, coughing etc.)	Acoustic	10	Lab	LOSO, LOPO	F_1: 79.5% (SD), 49.6% (SI)

Table 13: Summary of Classification Performance for ADM Systems Continued

Ref.	Objective	Sensor(s)	# of Subj	Data Source	Cross Validation	Result (Acc, F_1, TPR, FPR)
Kawano [44]	50-class food image classification	Image		First-person	5-fold	57.5%
Li [12]	4-class classification (drinking, chewing, coughing, speaking)	Accelerometer	8	Lab	10-fold	Acc: 80.98%
Rahman [5]	12-class classification (eating, drinking, speaking, clearing throat, coughing etc.)	Acoustic	14	Lab	LOSO, LOPO	F_1: 86.6% (SD), 67.6 (SI)
Walker [50]	5-class classification (swallow solid, swallow liquid, swallow saliva, speech, others)	Acoustic	7	Lab	10-fold	TPR: 90 - 95%, FPR: 10 - 15%
Olubanjo [30]	5-class classification (swallowing, chewing, speaking coughing, clearing throat)	Acoustic	5	Lab	Hold-out	F_1: 87.4% (SD)
Anthimopoulos [87]	11-class food image classification	Image	n/a	Online	5-fold	Acc: 77.6%
Bi [6]	7-class food classification (apple, carrot, chips, cookie, peanut, walnut, water)	Acoustic	12	Lab	4-fold	Acc: 87.1% (SD), 84.9% (SI)
Bettadapura [91]	5-class food image classification	Image	n/a	First-person	3-fold	Acc: 63.3%

classified using SVM on color and texture features extracted from the segmented food regions. Finally, intake volume was calculated from before and after meal images using camera parameter estimation and model reconstruction methods. Classification of 19 food items was undertaken from a total of 63 first-person meal images. They obtained performance in the range of 84.2% to 95.8% depending on the training data size. Additionally, an estimation of food mass from volume estimation results was undertaken for 2 foods, garlic bread and yellow cake. A percentage error rate of 25.8% was obtained. Potential drawbacks of this work are: i) all images were acquired in the same room with the same lighting condition, ii) a calibration fiducial marker consisting of a color checkerboard was required in the camera field of view for geometry and color correctin of food images.

- *Yatani and Truong [7]*: The authors developed and presented *BodyScope*, a neckworn acoustic-based system for activity recognition. The system consists of a bluetooth headset, microphone and stethoscope chestpiece to amplify throat sounds. Their dataset included 10 samples of 12 activities (seating, breathing, eating cookies, eating bread, drinking, drinking with a sip, speaking, whispering, whistling, laughing, sighing and coughing) from 10 participants ranging in age from 20 - 30 years. Features from the time, frequency and cepstral domain were used with SVM, Naïve Bayes and 5-NN classifiers. They found SVM to provide the best classification results of 79.5% for subject-dependent classification and 49.6% for subject-independent classification. Potential drawbacks of this work are: i) all activities were recorded discretely and classification was not done on a continuous signal, ii) data was collected in a controlled lab study.

- *Rahman et al. [5]*: A neckworn system called *BodyBeat* was developed and presented for recognizing non-speech body sounds including eating. The system design includes a piezoelectric sensor-based microphone surrounded by soft and hard silicone layers in a 3D printed capsule for internal and external acoustic isolation, respectively. They compared a total of 7 microphone types to evaluate frequency response and susceptibility to various kinds of external noise (white, social, traffic and conversational noise). They found

the custom-made brass piezoelectric microphone type with silicone diaphragm material to be optimal for recording non-speech sounds. Their dataset includes a 15-mins recording from 14 participants conducting 12 different activities namely: eating cookies, apple, bread, banana, drinking water, taking deep breaths, clearing throat, coughing, sniffling, laughing, speaking and being silent. Using frame-level features and window-level statistical features in combination with a linear discriminant classifier, they achieved an overall subject-dependent F_1 score of 86.6% and subject-independent F_1 score of 63.4%. Specifically for classifying eating activities, they obtained a recall of 70.35% and precision of 73.29%. A potential drawback of this work is that the data was collected in a controlled lab study.

CHAPTER 3

UNDERSTANDING THE ACOUSTIC PROFILE OF FOOD INTAKE ACTIVITIES

A food intake cycle is a non-stationary, time-varying process that primarily includes bites, chews and swallow events. As seen in chapter 2.2, various sensor types have been used towards automatic food intake monitoring. However acoustic sensors, whether in single- or multi-modal units, are amongst the most common sensor types used towards food intake monitoring. Using an acoustic sensor, several factors can affect the output signal that is recorded including sensor location, microphone type, and/or food type being consumed. Solid foods can range in texture from hard crunchy to soft. In addition to food intake signals, an acoustic sensor will also record all other sounds within range including internally produced sounds and externally produced sounds (see Figure 14).

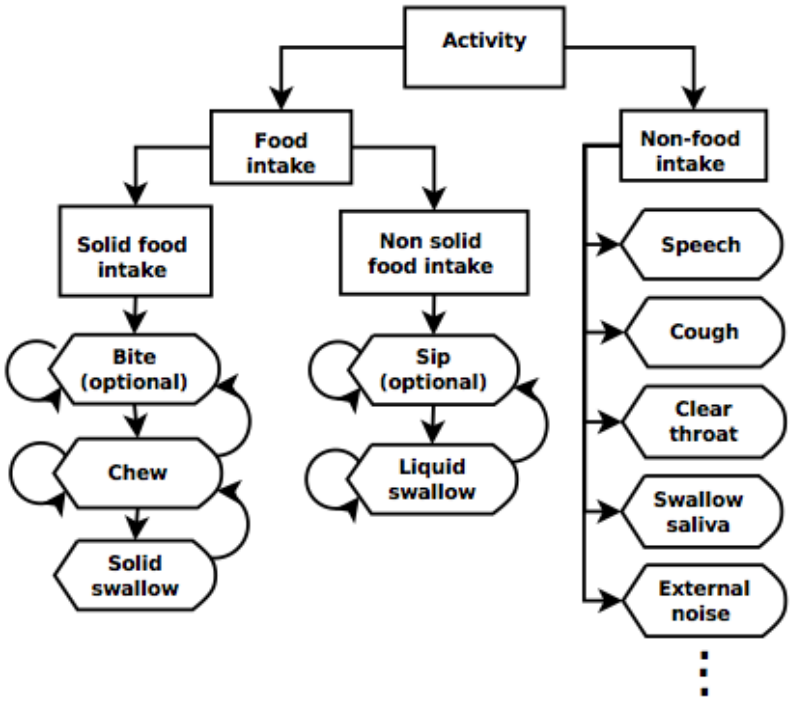


Figure 14: Activity breakdown for acoustic food intake monitoring systems

The objective of this chapter is to explore the acoustic signature of food intake events, with particular emphasis on chew events which are dominant during consumption of solid foods. Temporal characteristics and spectral characteristics of acoustic food intake events are presented to serve as a reference point for future work on automatic recognition and detection, particularly in realistic non-laboratory environments.

3.1 Temporal Characteristics

For detection and recognition of any event in a continuous signal, understanding basic temporal characteristics of the event of interest is imperative. Temporal characteristics such as event duration, frequency of occurrence, maximum amplitude etc. can play a key role in determining the right variables for the detection algorithm. In this study, we explored the average, maximum and minimum duration of chew events within food intake cycles for five food types. We did not include amplitude description in our analysis since signal amplitude can vary significantly based on the recording location, microphone type and/or any amplification applied to the recorded signal. We focused on chews because they are predominant events during eating. In addition, previous work [79] has explored the acoustic profile of swallow events. To evaluate the frequency of chew events, the chewing rate (chews/s) was calculated by counting the number of chew events from the beginning of a food intake cycle (immediately after the bite, if present) to the first swallow or 10s of the chewing sequence, whichever came last. MATLAB and Audacity (http://audacityteam.org) were used for visualization and audio analysis. The total number of audible and visibly recognizable chew events in a sequence divided by the time duration of the sequence yielded the reported chewing rate. Lastly, the rate of decrease of the energy of events in a food intake cycle was calculated from a linear regression of the energy profile. Previous work [3, 101] observed only the decline in energy amplitude of chew events during a food intake cycle, but not the rate of decline for different food types.

3.2 Spectral Characteristics

Figure 15 shows a snapshot of acoustic food intake signals and their associated spectrograms. Unlike speech, food intake events have no fundamental frequency and are not harmonic in nature, instead there is a spread of energy across a range of frequencies. In the frequency domain, three main descriptors were explored; namely, spectral slope, spectral roll-off and tonality. Spectral slope describes the amount of decrease in the spectral amplitude, and it is computed by linear regression. Spectral roll-off was defined as the frequency point under which 90% of the signal energy is contained. Spectral flatness, which is a measure of the noisiness of the spectrum, was computed as a ratio of the geometric mean to the arithmetic mean of the energy spectrum as shown in equation (1). The tonality coefficient was derived from the spectral flatness per equation (2). Tonality of noisy signals should be close to 0 whereas tonal signals should be close to 1.

$$SpectralFlatness_{dB} = 10\log\left(\frac{\left(\prod_{k\epsilon F_i} a(k)\right)^{1/k}}{\frac{1}{k}\sum_{k\epsilon F_i} a(k)}\right) \tag{1}$$

$$Tonality = min\left(\frac{SpectralFlatness_{dB}}{-60}, 1\right) \tag{2}$$

where $a(k)$ is the amplitude in k-th frequency band and F_i is a larger frequency band containing several k bands.

3.3 Data Collection

Tracheal data was recorded from 12 subjects (7 males, 5 females) at a sampling rate of 16-kHz with 16-bit resolution using an iASUS NT3 throat microphone placed over the suprasternal notch. Subject's ages ranged from 24-33 yrs (mean age: 29 yrs), weights ranged from 60.8–97.5 kg (mean weight: 75.1 kg), heights ranged from 157.5–185.4 cm (mean height: 172.7 cm) and body mass index (BMI) ranged from 21–33.7 (mean: 25.49).

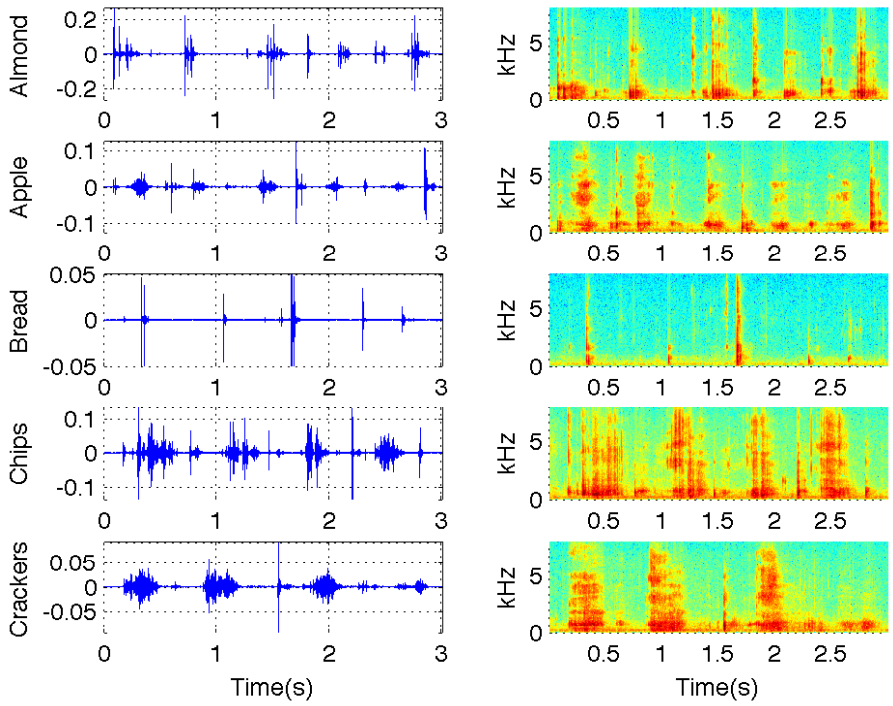
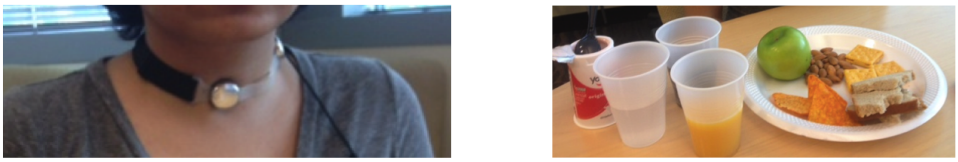


Figure 15: Acoustic food intake signals and associated spectrograms

It should be noted that BMI less than 18.5 is considered underweight, 18.5–24.9 is considered normal, 25–29.9 is considered overweight, while greater than 30 is considered obese [4]. The Institutional Review Board of Georgia Institute of Technology approved this study and all subjects signed a written consent form prior to the experiment. A LabVIEW program was set to automatically randomize the task order at the beginning of each data collection session. A total of 13 different tasks were included in this experiment: chewing and swallowing of solids (crackers, apple, almond nuts, chips and bread), swallowing of liquids (water, coke, yogurt, orange juice), as well as other tracheal and non-tracheal events (coughing, clearing the throat, speech and head motion). Figure 16 shows the experiment setup. The food selection was intended to represent various food textures/categories including wet crispy, dry crispy, hard crunchy, soft, etc., all of which impact the chewing and swallowing sounds. At least 39 tracheal events were collected and annotated per subject, and a minimum of 468 events for all 12 subjects.



(a) (b)

Figure 16: Food intake experiment (a) iASUS throat microphone (b) Various food for intake

3.4 Results

This section details the obtained results that describe the temporal and spectral profile in a food intake cycle.

3.4.1 Temporal Profile

We extracted temporal features from a total of 1471 manually annotated chew events (266 almond chews, 278 apple chews, 259 bread chews, 361 chips chews and 307 cracker chews). Table 14 shows the mean, maximum and minimum duration of chew events obtained in this work. We observed that chew events in the beginning of a chew sequence have longer durations than chew events towards the end of the sequence [102]. This is expected because at the beginning of a food intake cycle, the food is more solid and the chewing sound includes initial crushing of harder texture material. Of the five food types considered, crackers had the longest maximum, average, and minimum duration of 0.84 s, 0.27 ± 0.11 s and 0.06 s, respectively. Chewing of chips had the second longest maximum, average, and minimum duration of 0.70 s, 0.24 ± 0.09 s, and 0.04 s, respectively. Across all food textures considered in this work, the average duration of a chew event was 0.22 ± 0.1 s. In addition, the average chewing rate across all food textures was 1.64 chews/s, softer foods such as bread had a slower chewing rate of $\sim$ 1.37 chews/s, whereas dry crispy foods like chips had the fastest chewing rate of $\sim$1.81 chews/s.

Previous work [3] supports that there is a decline in the energy profile of events during an intake cycle. Our work takes this finding further by presenting the rate of decrease in the energy profile of food intake cycles. Using linear regression of the energy profile, we

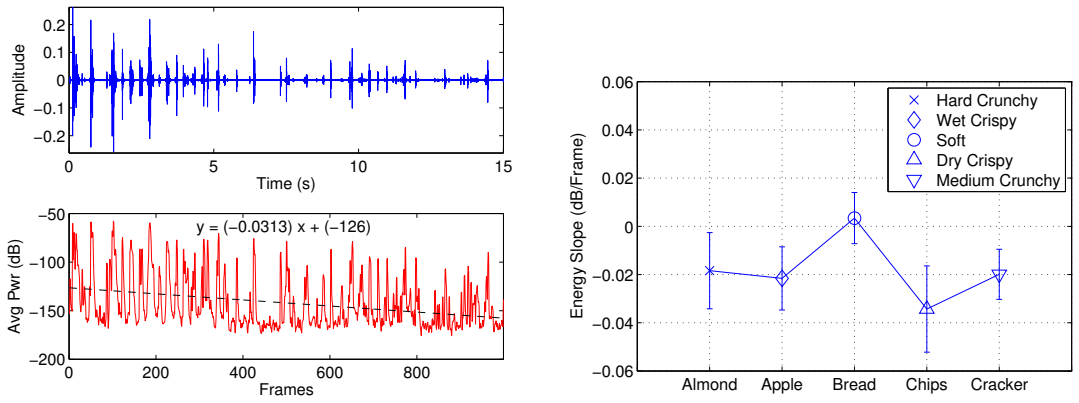
Table 14: Temporal profile for chew events

	Chew Events			Chew Rate (chews/s)
	Mean Dur. ± Std Dev. (s)	**Max. Dur. (s)**	**Min. Dur. (s)**	
Almond	0.17 ± 0.09	0.59	0.03	1.76
Apple	0.23 ± 0.11	0.67	**0.02**	1.67
Bread	0.17 ± 0.09	0.49	**0.02**	1.37
Chips	0.24 ± 0.09	0.70	0.04	1.81
Cracker	0.27 ± 0.11	**0.84**	0.06	1.58
Average	**0.22 ± 0.10**	**-**	**-**	**1.64**

observed the slope for food types of varying textures (see Figure 17). There was a more notable decrease in the energy profile, shown by the negative slope, of all food types except bread. In many cases, the energy profile for bread intake was slightly positive because chewing of soft food material had very low energy level and many times, swallowing of the bolus produced higher energy than was recorded from chewing. Overall, the mean slope of the energy profile across all food types considered in this study was $-0.0182\, dB/frame$, where the frame length was set to 30 ms. The greatest decline in the energy profile was observed for chewing of chips which is a dry crispy food type. It should be noted that food type did not have a statistically significant effect on the energy slope values obtained ($P > 0.05$).

3.4.2 Spectral Profile

Table 15 shows the spectral profile for all food types evaluated in this work. The highest average power was observed in chew events of almond intake, which was > 4 times the average power observed for chew events of bread intake. On average, 90% of the total energy in chew events was < 2 kHz. Chew events of soft food types like bread had the highest spectral roll-off point of 3.04 ± 0.93 kHz whereas harder foods like almonds, crackers, and chips had a spectral roll-off frequency of 1.31 ± 0.61 kHz, 1.64 ± 0.78 kHz and 1.71 ± 0.70 kHz, respectively. On average, the spectral slope across food types was $-5.4\, dB/kHz$. The steepest spectral slope was observed for chew events from almond intake and the least



(a) (b)

Figure 17: Energy profile during food intake cycle: (a) Sample energy slope graph (b) Energy slope across varying food textures

Table 15: Spectral profile for chew events

	Avg. Pwr	**Spectral Rolloff**	**Spectral Slope**	**Tonality**	
	(dB)	**(kHz)**	**(dB/kHz)**	**0 - 1.5 kHz**	**1.5 - 4 kHz**
Almond	-39.64 ± 3.58	1.31 ± 0.61	-6.2	0.36	0.25
Apple	-47.14 ± 3.46	2.15 ± 0.90	-5.3	0.33	0.25
Bread	-52.52 ± 4.81	3.04 ± 0.93	-4.5	0.30	0.27
Chips	-40.76 ± 3.07	1.71± 0.70	-5.5	0.31	0.23
Crackers	-46.01 ± 4.91	1.64 ± 0.78	-5.6	0.32	0.23
Average	**-45.21 ± 3.97**	**1.97 ± 0.78**	**-5.4**	**0.32**	**0.25**

steep spectral slope was observed for bread intake. This finding is consistent with the fact that there is more energy in chew events from almond intake than chew events from bread intake. From the tonality analysis, we observed that the tonality coeffecients across all food types was < 0.36 and had a small variance. This supports that chew events of varying food types can be characterized as having a relatively flat and noisy spectrum.

3.5 Discussion

Chew events from eating almonds had the highest average power, steepest spectral slope and smallest spectral roll-off frequency. Chewing of apple had an average power of less than half the average power from chewing almonds and chips. Chewing of bread had the

lowest average power of less than one-fourth the average power of chewing almonds and had the highest spectral roll-off frequency. Chew events from eating chips had the highest chewing rate and relatively high mean duration, average power, and a steep spectral slope, whereas chewing of crackers had the highest mean event duration. These nuances can be used in classifying different food textures. For all food types in this study, the mean duration obtained was 0.22 s, the mean chewing rate was 1.64 chews/s, on average 90% of chew events energy was under 2 kHz and the mean spectral slope was $-5.4\,dB/kHz$. These more general characteristics provide prior knowledge that can be used towards detecting chew events in a noisy signal recording.

CHAPTER 4

TRACHEAL ACTIVITY CLASSIFICATION AND REAL-TIME SWALLOWING DETECTION

This chapter explores two main goals: 1) tracheal activity recognition to classify and identify food intake activities, chewing and swallowing, from amongst other common activities that can be recorded by an acoustic system in daily living, 2) real-time swallowing detection based on computationally inexpensive features.

4.1 Tracheal Activity Recognition Based on Acoustic Signals

Tracheal activity recognition can play an important role in continuous health monitoring for wearable systems and facilitate the advancement of personalized healthcare. Neck-worn systems provide access to a unique set of health-related data that other wearable devices simply cannot obtain. Activities including breathing, chewing, clearing the throat, coughing, swallowing, speaking and even heartbeat can be recorded from around the neck.

In this work [30], we explored tracheal activity recognition using a combination of promising acoustic features from related work and applied simplistic classifiers including K-Nearest Neighbor (K-NN) and Naive Bayes. For wearable systems in which low power consumption is of primary concern, we showed that with a sub-optimal sampling rate of 16-kHz, we achieved average classification results in the range of 86.6% to 87.4% using 1-NN, 3-NN, 5-NN and Naive Bayes. All classifiers obtained the highest recognition rate in the range of 97.2% to 99.4% for speech classification. This is promising to mitigate privacy concerns associated with wearable systems interfering with the user's conversations.

4.1.1 Data Collection

Tracheal acoustics was recorded from five healthy subjects (2 males and 3 females, ages 20-35 years). From the tracheal activities of interest, swallowing has a bandwidth up to 1.5-kHz, speech, up to 4-kHz [7], and chewing, up to $\sim$ 6-kHz depending on the substance

being chewed. On the other hand, coughing and clearing the throat can reach frequencies up to ~ 15-kHz. According to the Nyquist theorem, 16-kHz is sufficient to preserve important characteristics of chewing, swallowing and speech but possibly not all characteristics of coughing and clearing the throat [7, 29].

To account for physiological variations, experimental data was collected from subjects on two different days. Data from day 1 was used for training while data from day 2 was used for testing. Table 16 shows a summary of the dataset used in this experiment. The 'chewing' activity consisted of each subject chewing two crackers while the 'swallow' activity consisted of each subject swallowing some water and a few dry swallows when they were audible and visibly recognizable during activity labeling. The speech activity consisted of subjects reading the same text.

4.1.2 Feature Extraction

Tracheal activities were isolated and annotated from the continuous recording by listening to the audio stream, visually inspecting the signal, and validating the event label with the experimental procedure. Acoustic features were extracted from each isolated activity using a window size of 1000 samples (62.5 ms) with 50% overlap. In an effort to achieve good clustering of features from each activity, we compiled promising features that have been used to obtain acceptable performance from related works [7, 28, 29, 103]. A total of 47 features from the time, frequency, and cepstral domains, as shown in Table 17, were used for training and testing of each classifier. Features from each window frame per tracheal

Table 16: Tracheal activity classification: Data summary for five subjects

Activity	Day 1 - Training	Day 2 - Testing	Total Number
Chew	65	61	126
Clearing throat	51	48	99
Cough	51	52	103
Swallow	51	55	106
Speech	121	95	216
Total number of tracheal activities			650

event were then averaged to obtain one real number to represent each activity.

Table 17: Tracheal activity classification features

Time domain	Windowed energy [29], total variation, zero crossing rate [7]
Frequency domain	Power spectral density (PSD), spectral centroid, spectral roll-off [7]
Time-frequency domain	Discrete wavelet transform [28, 29]
Cepstral domain	Mel frequency cepstral coefficients (MFCCs) [103]

4.1.3 Classifier

K-NN and Naive Bayes classifiers were used in this work. Both classifiers were implemented in MATLAB using the Statistical Pattern Recognition Toolbox. Euclidean distance was used to determine the nearest neighbors for K-NN while normal distribution was used for the Naive Bayes classifier. The chosen value of K governs the degree of smoothing; hence, there is an optimum choice for K that is neither too large nor too small. For this reason, odd-number values of K ranging from 1 to 5 were explored and compared.

4.1.4 Results and Discussion

Standard information retrieval statistics was used to evaluate performance of the proposed tracheal activity recognition algorithm. A confusion matrix was used to evaluate performance of each classifier on subject-dependent bases. Performance metrics, *recall*, *precision* and F_1 score were calculated for each activity to compare classifier performance:

$$Recall = \frac{TP}{TP + FN};\ Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

A recall performance of 1 means that the event of interest was correctly classified on all occasions while a precision performance of 1 means that there were no false positives.

Recall and precision were calculated per equation (3). The best possible F_1 score is 1.

In K-NN classifier, each new data point is assigned to the class having the largest number of representatives from the K nearest points in the training dataset. Therefore, to avoid a tie situation in the majority voting scheme, we focused on odd-number values of K. Classifier results for each activity using K-NN, K = 1, 3 and 5, and Naive Bayes are shown in Figure 18. The 1-NN classifier achieved the highest F_1 score of 0.912 and 0.902 for chewing and swallowing classification respectively.

The 3-NN classifier achieved the highest F_1 score of 0.872 for classification of clearing the throat while 5-NN achieved the highest F_1 score of 0.754 for classification of coughing. Our confusion matrix showed that coughs were sometimes misclasssified as clearing the throat, a similar high energy activity. Naive Bayes classifier achieved the highest F_1 score of 0.994 for classification of speech. All classifiers in this study achieved the highest recognition rate for classification of speech. Since the mel frequency scale is a variant of the critical band scale, which is based on perceptual studies [103], we expect that having MFCCs as a feature for classification contributed to this high classification performance for speech. The ability to classify speech with almost perfect accuracy can mitigate privacy

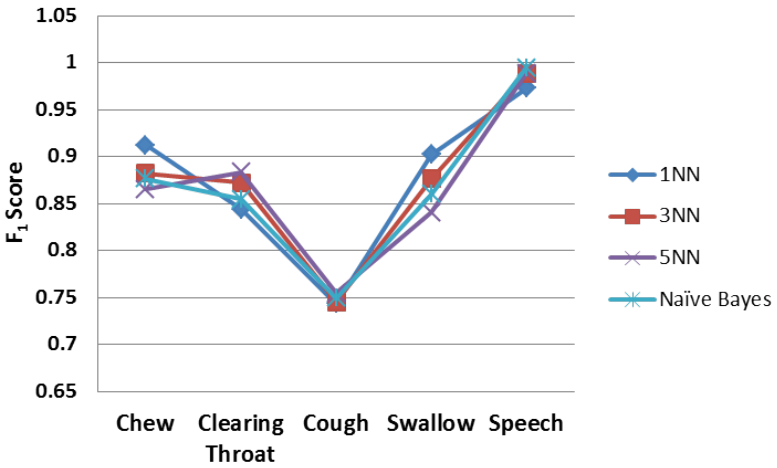


Figure 18: Tracheal activity classification - F_1 scores for 1-NN, 3-NN, 5-NN and Naive Bayes classifiers

Table 18: Tracheal activity classification: Summary of classifier performance

	1-NN	3-NN	5-NN	Naive Bayes
This Work [30]	0.874	0.873	0.866	0.867
Yatani, 2012 [7]	-	-	0.752	0.722

concerns associated with audio-based wearable health monitoring system by ensuring that the user's conversations can be eliminated before access is provided for further analysis on tracheal events of interest for health monitoring purposes.

Table 18 shows a summary of classifier performance for all tracheal activities considered. Each classifier's average performance ranged from 0.866 to 0.874. These classification results lead us to infer that although a sampling rate of 16-kHz is not sufficient to preserve all important characteristics in tracheal activities, it is sufficient to obtain good classification performance for tracheal activity recognition.

Comparing our results with the results presented by Yatani and Truong in [7] for tracheal activity recognition, our highest mean F_1 score of 0.874 was better than their highest F_1 score of 0.795 with support vector machine as the classifier. Using 5-NN, in this work we achieved an F_1 score of 0.866 while in [7] they achieved an score of 0.752 using leave-one-sample-per-participant out (see Table 18). Similarly, in this work we achieved an F_1 score of 0.867 using Naive Bayes while in [7] they achieved an F_1 score of 0.722 using the same classifier. These results lead us to infer that the list of features used in this study may be more effective than those used in [7]. It is important to note that the authors of [7] considered 12 activities in their study, which is more extensive than the activities considered here and therefore a limitation of this study. In addition, it should be noted that the results presented in [7] are the results used directly for comparison. Future work will include running the features presented in [7] on our dataset for a more fair comparison.

4.2 Real-Time Swallowing Detection Based on Tracheal Acoustics

The ability to automatically detect swallowing in real-time can provide valuable insight into eating behavior, medication adherence monitoring, and diagnosis and evaluation of swallowing disorders. In this work [29], we have developed a preliminary real-time swallowing detection algorithm based on acoustic signals that combines computationally inexpensive features to achieve comparable performance with previously proposed methods using acoustic and non-acoustic data. With a dataset that includes tracheal events such as speaking, chewing, coughing, clearing the throat, and swallowing of different liquids, our results show an overall recall of 0.799 and precision of 0.676.

Our work on real-time swallowing detection also has the potential to be used as a camera trigger for image dependent food intake monitoring systems; when the frequency of swallows increases, it may be assumed that the user is either eating or drinking something. This can save image storage space, reduce processing efforts on retrieved images and privacy concerns associated with taking pictures at fixed time intervals throughout an entire day.

4.2.1 Data Collection

Acoustic data was collected using the iASUS NT3 throat microphone placed over the suprasternal notch of the trachea with a sampling rate of 16-kHz (same dataset as mentioned in section 4.1.1). Figure 19 shows that according to the Nyquist theorem, a sampling frequency of 16-kHz is sufficient to preserve important characteristics of the swallowing sounds that reach a maximum frequency of 1.5-kHz [7]. Data was recorded from 5 subjects (2 males, 3 females, ages 20 - 35 years) with no history of swallowing disorders as they were instructed to perform a variety of activities. Recordings from two subjects were excluded from data analysis due to an incomplete experiment and the throat microphone not maintaining contact during one experiment. This study was approved by the Institutional Review Board of Georgia Institute of Technology and all participants signed a written consent prior to the experiment.

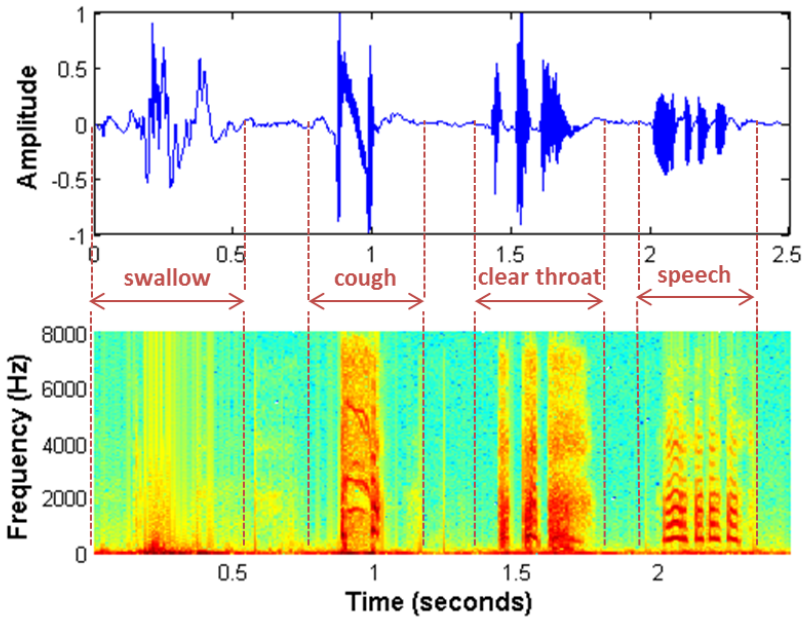


Figure 19: Spectrogram of common tracheal events

4.2.2 Methodology

The goal of an efficient real-time swallowing detection algorithm is to use minimal computational resources to achieve high swallowing detection accuracy. This implies the use of computationally inexpensive features to discriminate swallowing from other common tracheal sounds. To achieve real-time swallowing detection, we used 4 easy-to-compute features namely: windowed energy - equation (5), peak frequency - equation (6), Shannon entropy - equation (7) and wavelet entropy. The wavelet decomposition was computed using Coiflet 4 wavelet as in [28]. LabVIEW's Advanced Signal Processing Toolkit was used to extract coefficients of the discrete wavelet transform. The wavelet delta coefficients at decomposition level 3 were then converted into a scalar feature using Shannon's entropy.

Windowed energy (W.E) was described as:

$$W.E(X) = \sum_{i=0}^{n} \left| X_t(i)^2 \right| \tag{5}$$

where $X_t(i)$ is the discrete sample amplitude at time t, and n is the number of samples per window frame.

Peak frequency (P.F) is described as the frequency that has the highest energy:

$$P.F(X) = argmax_{[f=0,f_{max}]} |F_M(f)|^2 \tag{6}$$

where f_{max} is the highest available frequency in the signal and F_M represents the Fourier transform of the signal.

Shannon entropy (S.E) is described as:

$$S.E(X) = \sum_{i=0}^{n} X_t^2(i) \times log\left(X_t^2(i)\right) \tag{7}$$

where $X_t(i)$ is the discrete sample amplitude at time t, and n is the number of samples per window frame.

The acoustic data was processed and features were calculated with a non-overlapping 500ms window frame. A swallow event was detected when the pre-selected features per window frame was within the subject-dependent threshold range (R):

$$if\ [W.E(X) \in R_{W.E} \cap P.F(X) \in R_{P.F} \cap S.E(X) \in R_{S.E} \cap W.S.E(X) \in R_{W.S.E}]$$
$$\begin{cases} O_1(X) = 1; \\ otherwise,\ O_1(X) = 0 \end{cases} \tag{8}$$

where $R_{W.E}$, $R_{P.F}$, $R_{S.E}$, $R_{W.S.E}$ represent the subject-dependent threshold ranges for windowed energy, peak frequency, Shannon entropy and wavelet Shannon entropy respectively. The algorithm's output is represented by $O_1(X)$.

4.2.3 Results

The following definitions were used to analyze our results: true positives (TP) represents correctly detected swallows, false negatives (FN) represents undetected swallows and false positives (FP) represents incorrectly detected swallows. True negatives (TN) are ill-defined because they consist of times where a swallow does not occur and the algorithm does not detect one. We used standard measures for information retrieval to assess performance:

Selection of an appropriate subject-dependent threshold range was critical for our proposed real-time swallowing detection algorithm. A swallow was considered correctly detected if there was at least one detected swallowing event with a temporal center situated between time stamps of an annotated swallowing event, or if the temporal center of an annotated swallowing event laid in between the time stamps of at least one detected swallowing event.

Recall and precision performances were calculated for all subjects using equation (3). Table 19 shows a summary of the achieved results for each subject from our experiment. On average, our real-time swallowing detection algorithm achieved 0.799 recall and 0.676 precision. As shown in Table 20, our overall results are comparable with the best recall performance presented in [7] and [104]. We chose to compare our results with [7] and [104] because both of these papers include other tracheal sounds such as coughing, speaking, eating solid foods and drinking different liquids in their experiment. Since there is currently no agreed upon physiological explanation for what causes swallowing sounds [105, 106], there is also no direct association between descriptive features extracted and physiological happenings.

Table 19: Summary of real-time swallowing detection

	Experiment			
	Part I		Part II	
	Recall	**Precision**	**Recall**	**Precision**
Subject 1	0.867	0.62	0.88	0.73
Subject 2	0.727	0.50	0.65	0.81
Subject 3	0.824	0.67	0.65	0.60
Subject 4	0.944	0.59	0.85	0.889
Average	**0.84**	**0.59**	**0.76**	**0.76**

Table 20: Real-time results comparison with related work

	[29] Our Results	**[104] 2006**	**[7] 2012**
Recall	0.799	0.84	0.780
Precision	0.676	0.18	0.661
Real-Time	**Yes**	**No**	**No**

CHAPTER 5

INTAKE DETECTION IN NOISY ENVIRONMENTS

5.1 Source Separation for Target Enhancement of Food Intake Acoustics from Noisy Recordings

In this study, we explore the ability to learn spectral patterns of food intake acoustics from a clean signal and use this learned patterns for extracting the signal of interest from a noisy recording. Using standard metrics for evaluation of blind source separation, namely signal to distortion ratio and signal to interference ratio, we observed up to 20 dB improvement of separation quality in very low signal to noise ratio conditions. For more practical performance evaluation of food intake monitoring, we compared the detection accuracy for chew events on the mixed/noisy signal versus on the estimated/separated target signal. We observed up to 60% improvement in chew event detection accuracy for low signal to noise ratio conditions when using the estimated target signal compared to when using the mixed/noisy signal. Details of this collaboration work particularly with regard to the model built for source separation can be found in [31].

5.1.1 Results: Counts of chewing events

Detecting and counting of chew events in a food intake cycle is an objective metric that can be used to evaluate an automatic food intake monitoring system [2, 49]. Päßler and Fischer, in [49], presented and evaluated eight different algorithms for automated chew event detection on food intake sounds from consumption of six types of food. In this study, we apply the most successful and efficient algorithm from [49], maximum sound energy algorithm, for evaluation of the proposed source separation method presented in [31]. As with the maximum sound energy algorithm in [49], chew events were detected from a food intake cycle when the signal energy in a 23 ms frame segment and the following 12 frames exceeded a minimum threshold. Our minimum threshold value was found by comparing results of the chew event detection algorithm to a manually annotated ground

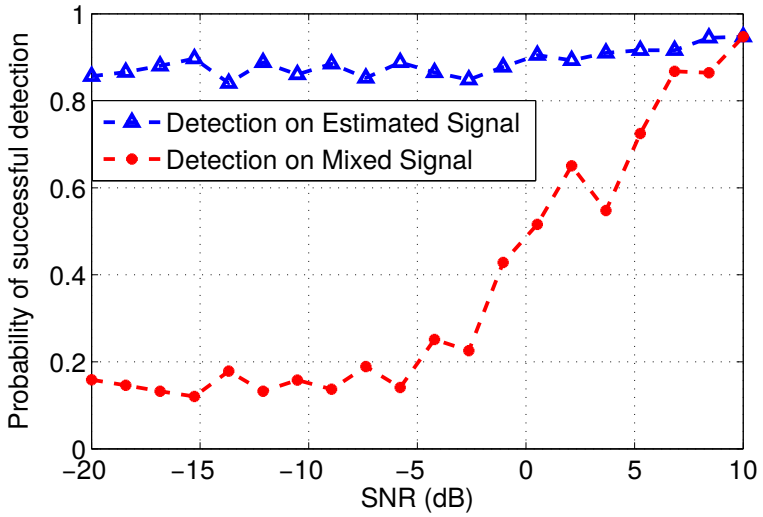


Figure 20: Chew event detection on mixed signal and estimated target signal relative to performance on clean signal, for various signal to noise ratios.

truth of the test signal to obtain the best possible performance. See the work presented in [49] for additional details on the *Maximum Sound Energy* algorithm for chew event detection. Performance of the maximum sound energy algorithm for chew event detection on the mixed signal and the estimated target signal, relative the clean signal, was then computed for various SNR values.

Figure 20 shows the results achieved from comparing chew event detection on the estimated signal with chew event detection on the clean signal. We observe that in negative SNR cases, when the noise signal completely overpowers the target signal, for example: $[-20\ -5]$ dB, there is $\sim$ 60% increase in chew event detection accuracy achieved from using the estimated signal. On the other hand, there is a little-to-no notable difference in detection accuracy when the SNR is $\geq$ 7dB. This result shows that in food intake monitoring applications, where the target is a low energy signal compared to the surrounding noise, in a loud restuarant for example, a huge benefit can be achieved from applying an intelligent source separation technique to estimate the clean signal compared to simply using the mixed signal for processing.

5.1.2 Discussion

In this study, we used recent source separation models and methods to denoise signal of interest in real-world single-sensor food intake acoustic data. Using only a limited recording, 1 minute, of the target signal, obtained in a silent laboratory setting, we showed that we can learn an adequate signal model for use in isolating the food intake acoustics from adverse background noise. We also showed the benefit of using this technique to exploit the denoised data for automatic monitoring applications is very high, compared to using the original mixture data. Additionally, in the case of automatic food intake recognition, we observed that using the proposed method to obtain an estimated target signal provided up to 60% improvement in chew event detection compared to the detection accuracy achieved on the mixed signal.

5.2 Detecting Food Intake Acoustics in Noisy Recordings using Template Matching

The objective of this work [32] is to explore detecting food intake events (particularly chew events) in the presence of restaurant background noise. In this preliminary study, an exemplary signal of one randomly picked subject from a larger dataset was used. This database includes tracheal recordings from 12 subjects (7 males, 5 females, age range: 24-33 years) as they ate five foods with varying textures (almonds, apple, chips, crackers and bread). Three templates were extracted from a clean signal to represent the beginning, middle and end phase of a chewing sequence. Then, each template was used with sliding window correlation for subject-dependent chew event detection on an independent mixed/noisy recording. The noisy signal was formed by instantaneous addition of a clean throat microphone recording during eating and restaurant noise recorded with the same throat microphone (see Figure 21). Results show that the template created from the end phase of a chewing sequence outperformed templates from the beginning and middle phases for detecting chew events in a continuous clean and noisy test signal. An F_1 score of 0.714 was achieved for

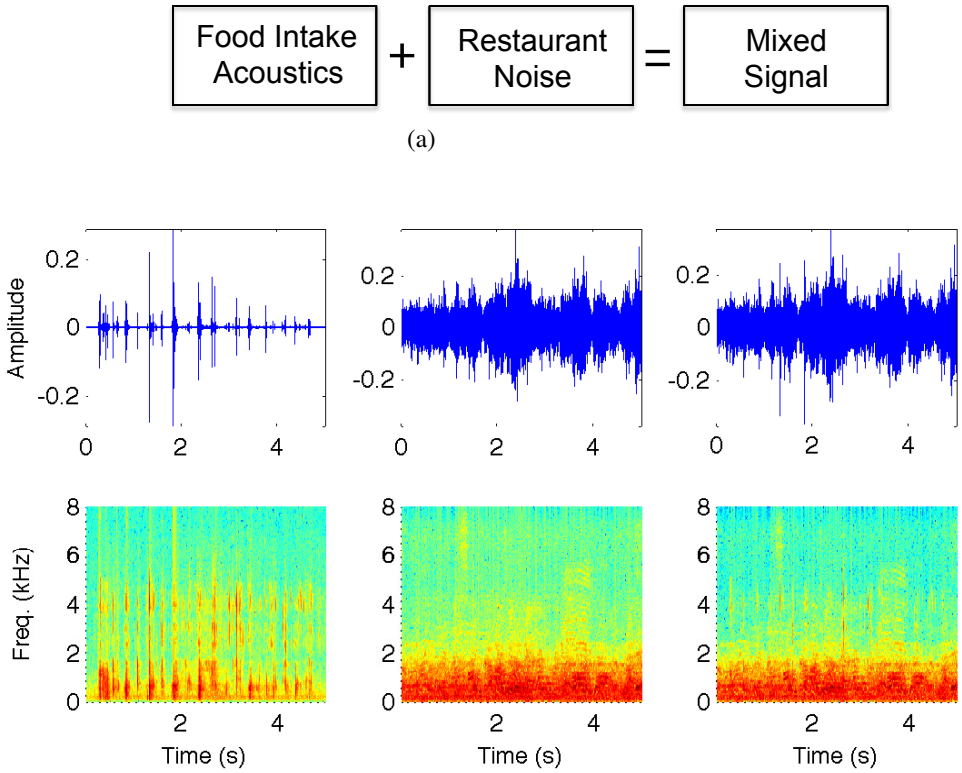


Figure 21: Artificially created noisy signal. a) Clean signal + Restaurant noise = Mixed/Noisy signal, b) Clean spectrogram, noise spectrogram, mixed/noisy signal spectrogram

detecting chews in very low signal-to-noise ratio of -10 dB.

5.2.1 Food intake templates

Food intake acoustics is primarily dominated by chew events. The chewing process involves gradual breaking down of the consumed food structure. This is evident by a decline in sound level in acoustic recordings [3, 60, 101]. In addition, previous research shows that chew events change during phases of a food intake cycle (from the time of putting a solid food item into the mouth to the time right before that item is swallowed) [3, 59, 60]. Päßler et al. hypothesize that a chewing cycle can be broken into three phases, namely: beginning, middle, and end [3]. The "beginning" phase is characterized by crushing sounds with high energy content at higher audible frequencies. The "middle" phase includes a mixture of crushing and grinding sounds, while the "end" phase is characterized by more wet and

smacking sounds with lower signal energy due to an increased amount of saliva in the mixture. Based on this hypothesis, in this work we develop three templates (T_1 , T_2 , T_3) for the three phases (beginning, middle, end) present in a food intake cycle, see Figure 22. Templates were developed based on the spectrogram of a clean chewing sequence by averaging the food intake sample windows in each phase (window length = 0.016 s (256 samples) with 50% overlap):

$$T_i \triangleq \frac{S_1 + S_2 + S_3 + ... + S_{N_s}}{N_s} \tag{9}$$

where S_1, ..., S_{N_s} is the food intake sample matrix $N \times L$ and N_s is the number of templates per phase. Each food intake sample matrix (S_k) is made up of the power spectral density values at each time-frequency unit. T_i is also a $N \times L$ matrix as it is the element-wise average of sample event matrices in each phase of the food intake cycle. Window length (L) of each food intake sample matrix was set to 0.112 s (1792 samples), about half of the average duration of a chew event activity. Using one template to represent each of the three phases of a food intake cycle, a dependence measure (correlation) was used to detect food intake acoustics in a continuous test signal. The detection accuracy obtained on a clean test signal is expected to be the best possible detection accuracy that can be achieved on the noisy test signal, saturated with multi-talker babble noise from a loud restaurant environment. A comparison between the detection accuracy with each template was also evaluated to identify whether partitioning/clustering strategy affects performance.

5.2.2 Detection Method: Sliding Window Correlation

Correlation is a common statistical measure that can be used to describe the relationship or dependence between two variables/objects. The hypothesis for application of the sliding window correlation method is this: the maximum correlation coefficient (ρ) between a given food intake template (T_i) and a sliding window (W_j) in the test signal should be high when W_j is part of or all of another chew event, and should be low when W_j is not part of

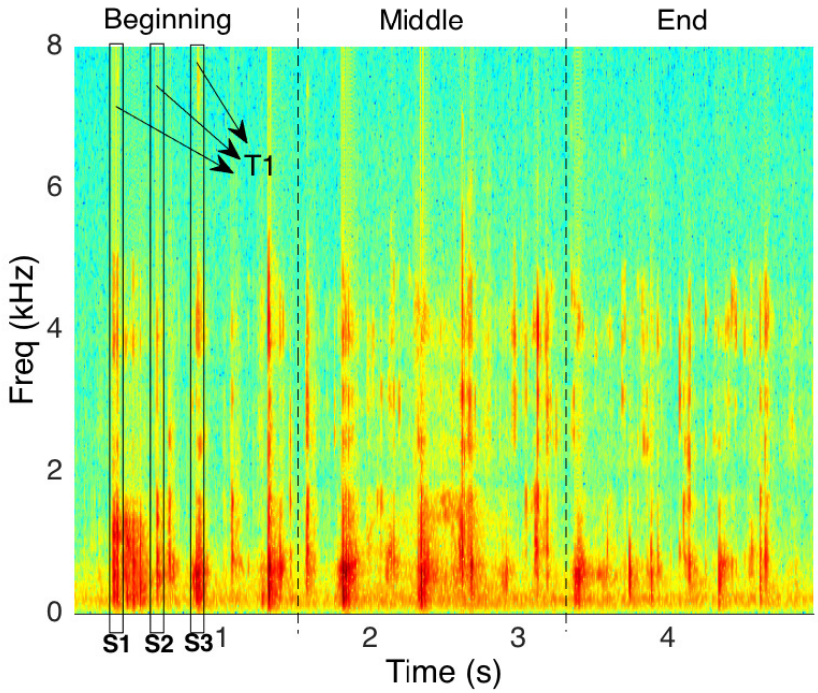


Figure 22: Template forming from clean food intake acoustics

a chew event:

$$\begin{cases} max\left[\rho\left(T_i, W_j\right)\right] \geq t_o; & W_j \in C_{d=1,..,n} \\ max\left[p\left(T_i, W_j\right)\right] < t_o; & W_j \notin C_{d=1,...,n} \end{cases} \tag{10}$$

where $\rho\left(T_i, W_j\right)$ is the normalized cross-correlation coefficient between food intake template (T_i) and test window (W_j), t_o is the optimum threshold for detection, and $C_{d=1,...,n}$ are the food intake events to be detected in the test signal.

All spectrograms were calculated using a frame length of 0.016 s (256 samples) with 50% overlap. The test signal's sliding window (W_j) dimension ($N \times L$) was chosen to be the same size as T_i to allow for calculation of normalized correlation coefficients. Matrices T_i and W_j, containing the power spectral density of each time-frequency unit on the spectrogram plot, were vectorized to enable use of the single dimension cross-correlation function:

Table 21: Template comparison for food intake detection

		T_1	T_2	T_3
Clean Signal	F_1 (%)	0.606	0.703	**0.833**
Noisy Signal	F_1 (%)	0.600	0.595	**0.714**

$$\rho\left(T_i, W_j\right)[n] = \sum_{-L}^{+L} \bar{T}_i[m] * W_j[m+n] \tag{11}$$

5.2.3 Results

Table 21 shows the detection performance of each template (T_1, T_2, T_3) used to represent the beginnng, middle and end phases of a food intake sequence. The best F_1 performance of 0.833 and 0.714 on the clean and noisy signals respectively was achieved using the end phase (T_3) for detection. An F_1 score of 0.703 was achieved using the middle phase template (T_2) compared to the F_1 score of 0.606 achieved using the beginning phase template (T_1) on the clean test signal. Meanwhile, almost the same performance of $\sim$ 0.6 was achieved with both the beginning and middle templates on the noisy signal. This could be an indication that a template built from the end phase (T_3) of a food intake cycle contains characteristics that are present in both the beginning and middle phases. Meanwhile, characteristics present in the beginning and middle phases are not always present in the end phase of a food intake cycle.

To compare our work with previous work on chew detection in a food intake cycle, we implemented the best computational inexpensive chew event detection algorithm (maximum sound energy) proposed in [49]. This algorithm detects a chew event in a chewing sequence when signal energy reaches a maximum greater than a certain threshold. See section IV-A of [49] for more details. Since this is a threshold based method, we evaluated threshold values in the range of [0.01, 1] to determine the best achievable detection performance on the clean signal then used this threshold for detection on the noisy signal (SNR = -10dB). Table 22 shows the performance results for our proposed template-matching

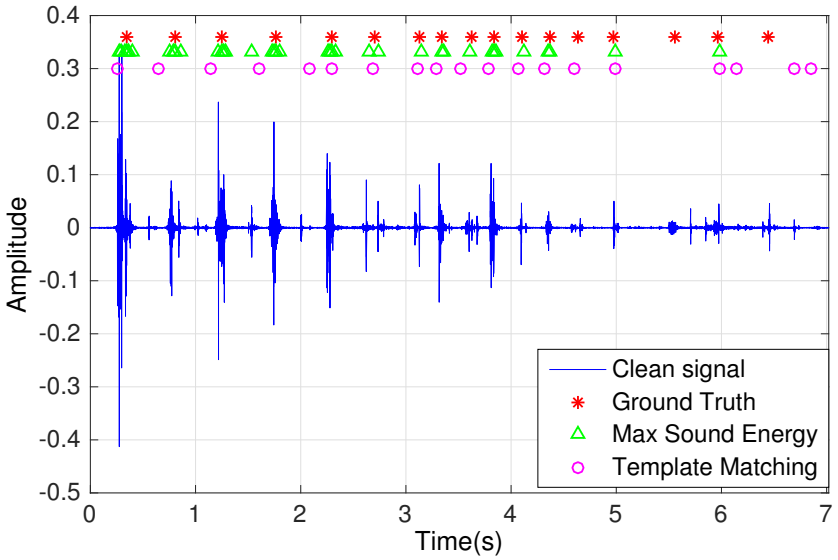
Table 22: Template matching comparison with related work

		Template-matching (This Work)	Max. Sound Energy [49]
Clean Signal	Rec.	0.882	0.824
	Pre.	0.790	0.933
	F_1	**0.833**	**0.875**
Noisy Signal (SNR=-10 dB)	Rec.	0.882	1
	Pre.	0.600	0.106
	F_1	**0.714**	**0.192**

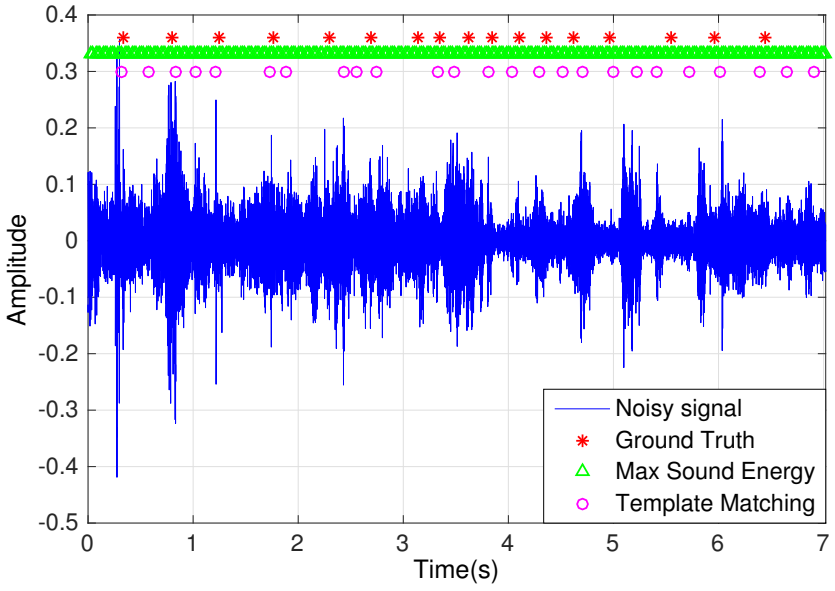
method compared with the "maximum sound energy" algorithm from [49]. All recall, precision and F_1 scores were calculated per equation (4).

The best F_1 score of 0.875 achieved with the maximum sound energy algorithm outperformed the best F_1 score of 0.833 achieved with our proposed template-matching algorithm on the clean signal. As expected, a degradation in performance was observed for chew event detection using both algorithms in a noisy signal with very low SNR of -10 dB. The F_1 score of the maximum sound energy algorithm dropped significantly to 0.192 while the template-matching algorithm maintained a more acceptable F_1 score of 0.714. The difference in detection performance in a noisy recording can be attributed to the fact that the maximum sound energy signal does not rely on a model or template that resembles the signal of interest while our template-matching algorithm works with known sample of the event of interest.

Figure 23 shows the detection performance of both the maximum sound energy algorithm from [49] and our proposed template-matching algorithm compared to the ground truth annotation on the clean and noisy signals respectively. For easy visualization of the detection performance only the center point of each annotation is shown. As can be observed from Figure 23b, the maximum sound energy algorithm practically detects every frame as a chew event in the noisy signal, this leads to very low precision of 0.106 and therefore overall low detection performance.



(a)



(b)

Figure 23: Chew detection a) Detection on clean signal b) Detection on noisy signal ($SNR = -10dB$)

5.2.4 Discussion

Acoustic-based systems for automatic food intake monitoring must have the ability to perform in various environmental conditions, including noisy environments. Results from this study show the potential for detecting food intake events in noise-saturated signals. This is useful for detecting periods of food intake in a continuous signal, chew count in food intake periods and eating rate. With the proposed template-matching method, F_1 score of 0.833 and 0.714 was achieved on a clean and noisy (-10 dB SNR) signal. The best detection performance was observed using a template built from the end phase of a chewing sequence compared to templates from the beginning and middle phases. This finding implies that the template used plays a notable role in detection performance. Future work includes comparing template-matching and hidden markov models for the same purpose of detection food intake events in noise-saturated signals.

CHAPTER 6

FUTURE RESEARCH RECOMENDATIONS FOR AUTOMATIC DIETARY MONITORING

This chapter begins with evaluating the acceptability of neckworn systems for dietary monitoring since this is the approach that was explored in our research work presented in chapters 4 and 5. After this, a performance benchmark of state-of-the-art ADM systems is presented, followed by research recommendations for future research based on the comprehensive review conducted.

6.1 Evaluating Acceptability of a Food Intake Neckwear System

A survey was conducted using written questionnaires after the food intake data collection described in section 3.3. The objective of this survey was to evaluate user acceptability criteria for a wearable dietary monitoring system.

6.1.1 Post-Experiment Questionnaire

Subjects were required to complete a post-experiment questionnaire. Figure 24a shows the questions included in the survey. All questions except Question 1 (Q1) were to be answered on a Likert rating scale (1-5), where a score of 1 equates to 'not willing/interested/comfortable' and a score of 5 equates to 'very willing/interested/comfortable'. Q1 required a yes/no response.

Out of the 12 subjects included in this study, only 3 subjects answered 'yes' to Q1 of the post- experiment questionnaire. Of those 3 subjects, 1-month was the maximum time that these subjects had consistently used a wearable health monitoring system. Additional input from subjects with longer prior experience using other wearable health monitoring systems is needed to better assess their views towards a wearable food intake monitoring system.

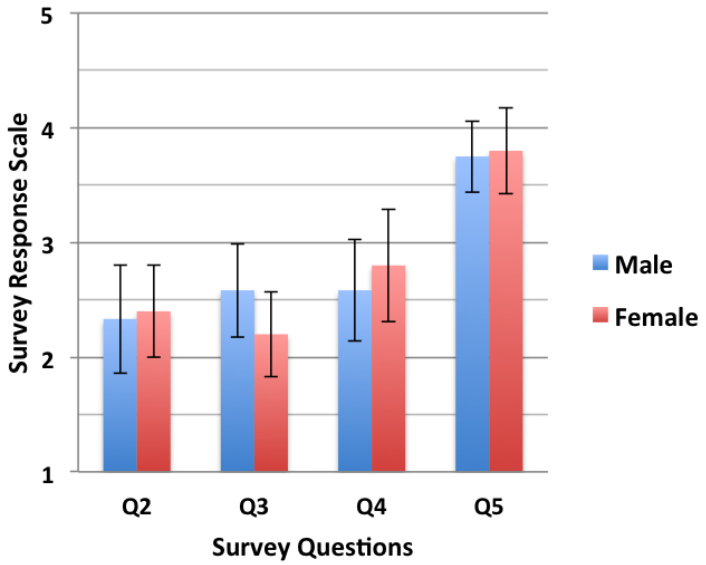
Figure 24b shows a summary of subject responses to post-experiment questionnaire Q2

to Q5. The standard error for each question was included on the bar graph to show how subject responses varied from the mean value. Standard error of a sample (also known as the standard deviation of the sample mean) is defined as follows: $S_e = \frac{\sigma}{\sqrt{n}}$, where σ is the sample's standard deviation and n is the sample size, which in our case was 12 subjects.

As can be seen in Figure 24, there were no major differences between the responses of male versus female subjects on the post-experiment questionnaire. Of the post-experiment questions Q2 – Q5, Q5 received highest average rating, meaning that subjects were fairly comfortable with the throat microphone used for recording during meal consumption and

1. Have you ever used any wearable health monitoring system such as Jawbone, Fitbit, Pedometer etc.? If yes, how long have you used it?
2. How willing are you to use a neckwear system that can monitor your health and daily activities?
3. How willing are you to use several wearable systems that work together to monitor your health and daily activities?
4. How interested are you in a wearable system that can monitor your daily food intake?
5. How comfortable was the throat microphone during meal consumption?

(a)



(b)

Figure 24: Post-experiment questionnaire towards a food intake neckwear system (a) Survey questions, (b) Survey responses

their eating experience was not impeded by the neckwear system. On the other hand, Q2 and Q3 received the lowest ratings. In response to Q2, subjects were not eager about the idea of a neckwear system for health monitoring. This may be because neckwear systems can be relatively obtrusive and uncomfortable when secured on too tight on the neck. Responses to Q3 suggest that subjects are not eager to use several wearable systems for health monitoring purposes. This supports the idea that users would prefer to use one wearable system instead of several wearable systems for health monitoring. The average response to Q4 was slightly above half of the rating scale, indicating subjects are favorably disposed to a wearable system that can monitor daily food intake.

6.2 Future Research Considerations

High-level tasks (see Figure 5) may require real-time processing while long-term behavioral monitoring and trend analysis can be done offline. As with all wearable systems, important criteria for acceptability and usability of an ADM system are that it is: portable and lightweight, robust, energy-efficient, minimally obtrusive, privacy-preserving, flexible to support new users, inexpensive and aesthetically pleasing. An added challenge relevant for the design of activity recognition systems is the tradeoff between accuracy, system latency and processing power [107]. Similar to the challenges highlighted in [78], ADM systems should be robust to intraclass variability, interclass similarity, null class problem and class imbalance. Below is a list of future research considerations for ADM systems:

- *Single-unit, multi-modal system*: Common sensors used towards dietary monitoring have obvious advantages and drawbacks, therefore it is envisioned that a multi-modal system is needed to fully tackle the problem of continuous monitoring. To enable a portable form factor, a single-unit multi-modal system is preferred and recommended over a multi-unit multi-modal system. Recognition of a broader set of activities and estimation of more dietary parameters in diverse environmental conditions is feasible when multi-modal sensors are selected carefully. Low-power sensors can be used for

gating or triggering higher-power, more detail capturing sensors during eating moments. For example, inertial sensors use less power than acoustic and image sensors. Amongst these, accelerometers use approximately one-tenth the power of gyroscopes and have been shown capable of detecting eating periods in [51]. In addition, inertial sensors can be considered as privacy preserving because a user cannot be easily identified from the inertial dataset. Therefore, such a sensor is the preferred sensing modality for continuous recording and detection to activate other sensors.

- *Recognition and Evaluation dataset*: To enable development of robust ADM systems with repeatable signal analysis methods, it is important to have a large, open-access, comprehensive, naturalistic and multi-day dataset for building and testing recognition models. Quantitative comparison of developed ADM algorithms is currently limited by the fact that each system works with a different dataset that is sometimes biased to the dietary activity of interest. A few public datasets relevant for dietary monitoring include kitchen and food preparation datasets [108–110], the Pittsburgh Fast-food Image Dataset (PFID) [100], iEatSet [41] and iHEARu-EAT [111]. Most of these datasets focus primarily on image, video and/or inertial sensing of particular events of interest for development of classification algorithms. It is envisioned that a multimodal dataset that includes acoustic data in addition to the aforementioned sensing data in a long-term, naturalistic recording environment will facilitate further research work.
- *Hierarchical structure*: Energy efficiency of wearable ADM systems is crucial to maximize the limited battery-life before need for a recharge. Hierarchical structures can reduce computational overhead and improve privacy-preservation by triggering the detail-capturing sensors (e.g. image and acoustic) and low-level classification algorithm less often. High-level food intake detection in free living conditions can be implemented as a first step by privacy-preserving sensors (e.g. inertial) and used to limit further processing to only detected eating periods. Dong et al. [51] observed

22.4 h of eating out of 449 total hours of free-living recording, which is a ratio of 1 to 20 for eating versus non-eating class. Therefore, a hierarchical approach that facilitates low-level processing solely on relevant data segments can lead to significant power saving and improved privacy.

- *Semi-supervised annotation and learning methods*: Ground truth annotation is a very expensive and tedious task especially for long-term activity recognition. Therefore, it is important to consider future implementation of semi-automatic annotation methods such as [112, 113], and semi-supervised learning methods such as co-training [114, 115], and weakly supervised learning [116]. These methods use a small labeled dataset to facilitate further labeling of a larger dataset, which is then used for recognition. Semi-supervised methods are particularly important for a free-living dataset which presents the added challenge of collecting reliable ground-truth annotation.
- *Context-aware design*: Previous activity recognition literature [117, 118] has shown the benefit of including prior probabilities for a given activity based on contextual information. Intelligent ontology models that define relationships and constraints among activities, artifacts, persons, communication routes and symbolic locations can enhance performance of solely statistical methods for activity recognition [118]. Work by Bettadapura et al., in [91], is amongst the few that have leveraged context for automatic food recognition. Without location information, food-category classification accuracy across 600 images was 15.7%, whereas when location prior was included for restaurant eating, their classification accuracy across 5 cuisines improved to 63.3% [91].
- *Concurrent activity evaluation*: Much of the research towards automatic dietary monitoring have been conducted on datasets with discrete activities happening in time. Whereas in real-world settings, people tend to eat while doing other activities such as talking, watching television, working, commuting etc. Therefore, concurrent (or

composite) activity evaluation is an important area for future work in this field. The recently published iHEARu-EAT dataset [111], which focuses on acoustic data from eating while speaking, is amongst the few publicly available datasets that presents concurrent data for eating recognition. Hu and Yang [119] propose the use of skip-chain conditional random fields (SCCRF) while Helaoui et al. [120] propose the use of Markov logic for recognizing interleaved and concurrent activities.

CHAPTER 7
CONCLUSION

In this thesis, we aimed to futher the state-of-the-art research on automatic food intake monitoring using wearabale sensor-based systems. Although this is still an open problem, we believe that we have made notable contribution to this field through the following accomplishments:

Task_1 Research results that present the acoustic profile (temporal and spectral profile) of chew events for various food types consumed by different subjects.

Task_2 An efficient tracheal activity recognition algorithm for acoustic-based dietary monitoring systems. The proposed algorithm achieved an F_1 score of $\sim 90\%$ for classifying chews and swallows from amongst other common tracheal activities.

Task_3 The first real-time swallowing detection algorithm for acoustic-based dietary monitoring systems. The proposed algorithm achieved an overall performance of 79.9% recall and 67.6% precision. This real-time swallowing detection work also has the potential to be used as a trigger for passive food intake monitoring systems that switch on more power consuming sensors such as a camera during automatically determined food intake periods.

Task_4 Algorithm for target enhancement of food intake acoustics from noisy recordings. The proposed algorithm uses learned spectral patterns of food intake acoustics from a clean signal to extract the signal of interest from a noisy recording. Up to 60% improvement in chew event detection accuracy was obtained when using the estimated target signal compared to using the raw noisy signal.

Task_5 Template matching algorithm for detection of food intake acoustic events in noisy recordings. The proposed algorithm achieved an F_1 score of 71.4% for detecting chew events in very low SNR signals of -10 dB compared to the F_1 score of 19.2% achieved by the maximum sound energy algorithm presented in a related work.

Task_6 A comprehensive literature review of methods for automatic food intake monitoring to identify the best performing and more promising approaches as well as to identify research gaps and motivate further work. A key recommendation for future consideration towards development of robust, wearable, sensor-based ADM systems is: design of a context-aware, single-unit multi-modal sytem with hierarchical structure signal analysis approach, capable of semi-supervised annotation and learning and robust dietary monitoring in various recording environments.

7.0.1 Limitations and Future Work

This thesis has covered several aspects of automatic dietary monitoring with a slightly higher focus on a neckworn acoustic-based system. Although this can be a reliable approach, there are limitations that can be better handled in future studies. First and foremost, using the neck as the sensing location of choice for tasks 1 - 5 above was not well studied in advance before carrying out further studies. Our post-experiment questionnaire presented in Section 6.1 showed somewhat unfavorable results (< 2.5 on a 5-point likert scale) towards a food intake neckwear system. As seen in Table 2, neckworn systems often requires close sensor contact to the user's neck. This can lead to a tight-fitting and hence uncomfortable system around the user's neck. From the other sensing locations presented in Table 2, the wrist is the least obtrusive wearable location and presumably a better location for high-level detection of food intake periods in daily living.

Task_1 A limitation of the acoustic profile of chew events study, discussed in full detail in chapter 3, is that some of the results obtained, such as energy slope and tonality, did not show statistically significant differences between the food types explored. The goal of this work was to observe the hypothetically supposed differences in acoustic signature of intake cycles for different food types/textures. This differences can then inform future work towards classification of solid food texture based on acoustic signals. However, parameters that showed no statistically significant differences should be further explored, in addition to other temporal, spectral and even cepstral

parameters in future work.

Task_2&3 A limitation of the tracheal activity recognition algorithm and real-time swallowing detection algorithm, discussed in full detail in chapter 4, is that only subject-dependent classification and detection results were analyzed and presented. To reduce amount of data needed for individual calibration, a subject-independent ADM system which can be adaptively trained to each new user is preferred. In addition, a larger database of tracheal activities, much lager than the one used in our work, should be utilized in future work to improve reliability and robustness of the developed system.

Task_4&5 A limitation of the algorithm for target enhancement of food intake acoustics and template matching algorithm, discussed in full detail in chapter 5, is that artifically generated noisy recordings (clean signal + noise signal) was used for testing. Although this choice was made to enable full knowledge of the ground truth labels from the clean signal, some studies do show that artifically generated noisy recordings can not serve as a full substitute for realistic noisy recording because both signals are not exactly equivalent [121]. In addition, our proposed noise handling algorithms were only tested on a small subject population. Future studies should include video recording to assist with ground truth annotation and utilize a larger database to improve reliability and robustness of the developed system.

Finally, future work should consider all recommendation and gaps presented in 6.2 to facilitate development of envisioned wearable system for automated dietary monitoring. Despite the aforementioned limitations of this work, we believe that our studies, developed algorithms and results contribute to efforts towards sensor-based dietary monitoring. In addition, we presented the first comprehensive review of literature on unobtrusive and wearable methods to identify the best approaches and pin-point gaps in the research.

REFERENCES

[1] O. Amft and G. Tröster, "On-Body Sensing Solutions for Automatic Dietary Monitoring," *Pervasive Computing, IEEE*, pp. 62–70, 2009.

[2] M. Shuzo, S. Komori, T. Takashima, G. Lopez, S. Tatsuta, S. Yanagimoto, S. Warisawa, J. Delaunay, and I. Yamada, "Wearable Eating Habit Sensing System Using Internal Body Sound," *J. of Adv. Mech. Design, Systems, & Manufacturing*, vol. 4, no. 1, pp. 158–166, 2010.

[3] S. Päßler, M. Wolff, and W. Fischer, "Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food.," *Physiol. Measurement*, vol. 33, pp. 1073–93, 2012.

[4] E. Sazonov, S. Schuckers, P. Lopez-Meyer, O. Makeyev, N. Sazonova, E. L. Melanson, and M. Neuman, "Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior.," *Physiological Measurement*, vol. 29, no. 5, pp. 525–41, 2008.

[5] T. Rahman, A. Adams, M. Zhang, E. Cherry, B. Zhou, H. Peng, and T. Choudhury, "BodyBeat: A mobile system for sensing non-speech body sounds," in *MobiSys'*, 2014.

[6] Y. Bi, M. Lv, C. Song, W. Xu, N. Guan, and W. Yi, "AutoDietary: A Wearable Acoustic Sensor System for Food Intake Recognition in Daily Life," *IEEE Sensors Journal*, vol. 16, no. 3, 2015.

[7] K. Yatani and K. Truong, "BodyScope: A wearable acoustic sensor for activity recognition," in *Ubiquitous Computing*, 2012.

[8] M. E. Rollo, S. Ash, P. Lyons-Wall, and A. Russell, "Trial of a mobile phone method for recording dietary intake in adults with type 2 diabetes: Evaluation and implications for future applications," *Journal of Telemedicine and Telecare*, vol. 17, no. 6, pp. 318–23, 2011.

[9] D. H. Wang, M. Kogashiwa, and S. Kira, "Development of a new instrument for evaluating individuals' dietary intakes," *Journal of the American Dietetic Association*, vol. 106, no. 10, pp. 1588–1593, 2006.

[10] C. K. Martin, J. B. Correa, H. Han, H. R. Allen, J. C. Rood, C. M. Champagne, B. K. Gunturk, and G. A. Bray, "Validity of the Remote Food Photography Method (RFPM) for estimating energy and nutrient intake in near real-time.," *Obesity*, vol. 20, no. 4, pp. 891–9, 2012.

[11] E. Thomaz, A. Parnami, I. Essa, and G. Abowd, "Feasibility of identifying eating moments from first-person images leveraging human computation," in *SenseCam*, 2013.

[12] C. Li, Y. Chen, W. Chen, P. Huang, and H. Chu, "Sensor-embedded teeth for oral activity recognition," in *ISWC*, p. 41, ACM, 2013.

[13] H. Kalantarian, N. Alshurafa, T. Le, and M. Sarrafzadeh, "Monitoring eating habits using a piezoelectric sensor-based necklace," *Computers in Bio. & Med.*, vol. 58, pp. 46–55, 2015.

[14] E. Sazonov and J. Fontana, "A sensor system for automatic detection of food intake through non-invasive monitoring of chewing.," *IEEE Sensors Journal*, vol. 12, no. 5, pp. 1340–1348, 2012.

[15] Y. Dong, A. Hoover, J. Scisco, and E. Muth, "A new method for measuring meal intake in humans via automated wrist motion tracking," *Applied Psychophysiology and Biofeedback*, vol. 37, pp. 205–15, 2012.

[16] M. Farooq, J. Fontana, and E. Sazonov, "A novel approach for food intake detection using electroglottography," *Physiol. Measurement*, vol. 35, no. 5, pp. 739–51, 2014.

[17] B. Dong and S. Biswass, "Wearable sensing for liquid intake monitoring via apnea detection in breathing signals," *Biomedical Engineering Letters*, vol. 4, no. 4, pp. 378–387, 2014.

[18] M. Sun, J. D. Fernstrom, W. Jia, S. a. Hackworth, N. Yao, Y. Li, C. Li, M. H. Fernstrom, and R. J. Sclabassi, "A Wearable Electronic System for Objective Dietary Assessment," *Journal of the American Dietetic Association*, vol. 110, no. 1, pp. 45–47, 2010.

[19] J. Liu, E. Johns, L. Atallah, C. Pettitt, B. Lo, G. Frost, and G. Yang, "An intelligent food-intake monitoring system using wearable sensors," in *Body Sensor Network*, pp. 154–160, 2012.

[20] A. Kandori, T. Yamamoto, Y. Sano, M. Oonuma, T. Miyashita, M. Murata, and S. Sakoda, "Simple Magnetic Swallowing Detection System," *IEEE Sensors J.*, vol. 12, pp. 805–11, 2012.

[21] J. Fontana, M. Farooq, and E. Sazonov, "Automatic ingestion monitor: a novel wearable device for monitoring of ingestive behavior," *IEEE Trans. on Biomedical Engineering*, vol. 61, no. 6, pp. 1772–9, 2014.

[22] "The power of prevention: chronic disease... the public health challenge of the 21st century," *Nat. Center for Chronic Disease Prevention & Health Promotion, CDC*, pp. 1–16, 2009.

[23] U.S. Dept. of Health & Human Services, "Overweight and obesity statistics," 2010.

[24] T. D. Wade, A. Keski-Rahkonen, and J. I. Hudson, "Epidemiology of Eating Disorders," in *Textbook of Psychiatric Epidemiology*, pp. 343–360, John Wiley & Sons, Ltd, 2011.

[25] C. Yang and Y. Hsu, "A review of accelerometry-based wearable motion detectors for physical activity monitoring.," *Sensors*, vol. 10, no. 8, pp. 7772–88, 2010.

[26] O. Lara and M. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Comm. Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[27] R. Troiano, "Translating accelerometer counts into energy expenditure: advancing the quest," *J. of Appl. Physiol.*, vol. 100, no. 4, pp. 1107–8, 2006.

[28] E. Sazonov, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. Melanson, and M. Neuman, "Automatic

detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior," *IEEE TBME*, vol. 57, no. 3, pp. 626–33, 2010.

[29] T. Olubanjo and M. Ghovanloo, "Real-time swallowing detection based on tracheal acoustics," in *IEEE ICASSP*, 2014.

[30] T. Olubanjo and M. Ghovanloo, "Tracheal activity recognition based on acoustic signals," in *IEEE Eng. in Med. and Bio. Conf.*, 2014.

[31] A. Liutkus, T. Olubanjo, E. Moore, and M. Ghovanloo, "Source separation for target enhancement of food intake acoustics from noisy recordings," in *IEEE WASPAA*, 2015.

[32] T. Olubanjo, E. Moore, and M. Ghovanloo, "Detecting food intake acoustic events in noisy recordings using template matching," in *Int. Conf. on Biomedical and Health Informatics (BHI)*, 2016.

[33] "Dietary Guidelines for Americans," *U.S. Dept. of Agriculture, U.S. Dept. of Health and Human Services*, pp. 1 – 112, 2010.

[34] R. K. Johnson, "Dietary intake–how do we measure what people are really eating?," *Obesity research*, vol. 10, pp. 63S–68S, nov 2002.

[35] J. Witschi, "Short-term dietary recall and recording methods," in *Nutritional Epidemiology*, pp. 52–68, Oxford University Press, 2nd. ed., 1990.

[36] L. E. Burke, J. Wang, and M. A. Sevick, "Self-monitoring in weight loss: a systematic review of the literature.," *Journal of the American Dietetic Association*, vol. 111, no. 1, pp. 92–102, 2011.

[37] G. Block, "A review of validations of dietary assessment methods," *American Journal of Epidemiology*, vol. 115, no. 2, pp. 492–505, 1982.

[38] A. Black, A. Prentics, G. Goldberg, S. Jebb, S. Bingham, M. Livingstone, and W. Coward, "Measurements of total energy expenditure provide insights into the validity of dietary measurements of energy intake," *J. of American Dietetic Assoc.*, vol. 93, no. 5, pp. 572–579, 1993.

[39] J. Speakmen, *Doubly labelled water: Theory and practice*. Springer Science & Business Media, 1997.

[40] H. Cheng, Z. Liu, Y. Zhao, G. Ye, and X. Sun, "Real world activity summary for senior home monitoring," *Multimedia Tools and Applications*, vol. 70, no. 1, pp. 177–197, 2014.

[41] V. D. Kakra, N. P. V. D. Aa, and L. P. J. J. Noldus, "A multimodal benchmark tool for automated eating behaviour recognition," in *Measuring Behavior*, 2014.

[42] K. H. Chang, S. Y. Liu, H. H. Chu, J. Y. J. Hsu, C. Chen, T. Y. Lin, C. Y. Chen, and P. Huang, "The diet-aware dining table: Observing dietary behaviors over a tabletop surface," in *Pervasive Computing*, vol. 3968, pp. 366–382, 2006.

[43] F. Zhu, A. Mariappan, C. J. Boushey, D. Kerr, K. D. Lutes, D. S. Ebert, and E. J. Delp, "Technology-assisted dietary assessment," in *Proceedings of SPIE*, 2008.

[44] Y. Kawano and K. Yanai, "Real-time mobile food recognition system," in *Comp. Vision and Pattern Recog. (CVPR)*, pp. 589–593, 2013.

[45] Y. L. Zheng, X. R. Ding, C. C. Y. Poon, B. P. L. Lo, H. Zhang, X. L. Zhou, G. Z. Yang, N. Zhao, and Y. T. Zhang, "Unobtrusive sensing and wearable devices for health informatics," *IEEE Trans. on Biomed. Eng.*, vol. 61, no. 5, pp. 1538–1554, 2014.

[46] T. Miyazaki, G. C. de Silva, and K. Aizawa, "Image-based calorie content estimation for dietary assessment," in *IEEE Int. Symposium on Multimedia*, pp. 363–368, 2011.

[47] H. Junker, O. Amft, P. Lukowicz, and G. Tröster, "Gesture spotting with body-worn inertial sensors to detect user activities," *Pattern Recognition*, vol. 41, pp. 2010–24, jun 2008.

[48] O. Amft, M. Kusserow, and G. Tröster, "Bite weight prediction from acoustic recognition of chewing," *IEEE Trans. on Biomed Eng.*, vol. 56, no. 6, pp. 1663–72, 2009.

[49] S. Päßler and W. Fischer, "Food intake monitoring: Automated chew event detection in chewing sounds," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 1, pp. 278–89, 2014.

[50] W. Walker and D. Bhatia, "Automatic ingestion detection for a health monitoring system," *IEEE JBHI*, vol. 18, no. 2, pp. 682–692, 2014.

[51] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, "Detecting periods of eating during free-living by tracking wrist motion," *IEEE J. of Biomed. & Health Informatics*, vol. 18, no. 4, pp. 1253–1260, 2014.

[52] D. Ferriday, M. L. Bosworth, S. Lai, N. Godinot, N. Martin, A. A. Martin, P. J. Rogers, and J. M. Brunstrom, "Effects of eating rate on satiety: A role for episodic memory?," *Physiology & behavior*, vol. 152, pp. 389–396, 2015.

[53] O. Amft, "A wearable earpad sensor for chewing monitoring," in *IEEE Sensors Conference*, pp. 222–7, 2010.

[54] E. Thomaz, C. Zhang, I. Essa, and G. D. Abowd, "Inferring meal eating activities in real world settings from ambient sounds: A feasibility study," in *Int. Conf. on Intelligent User Interfaces*, pp. 427–431, 2015.

[55] H. Lu, D. Frauendorfer, M. Rabbi, M. Mast, G. Chittaranjan, A. T. Campbell, D. Gatica-perez, and T. Choudhury, "StressSense: Detecting stress in unconstrained acoustic environments using smartphones," in *Ubiquitous Computing*, pp. 351–360, 2012.

[56] A. Yadollahi and Z. Moussavi, "Acoustic obstructive sleep apnea detection," in *IEEE Eng. in Med. & Bio. Soc.*, pp. 7110–7113, 2009.

[57] M. Rofouei, M. Sinclair, R. Bittner, T. Blank, N. Saw, G. DeJean, and J. Heffron, "A non-invasive wearable neck-cuff system for real-time sleep monitoring," *Body Sensor Networks*, pp. 156–161, 2011.

[58] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, "Accurate and Privacy Preserving Cough Sensing Using a Low-cost Microphone," *Ubiquitous Computing*, 2011.

[59] J. Nishimura and T. Kuroda, "Eating habits monitoring using wireless wearable in-ear microphone," in *Int. Symposium on Wireless Pervasive Computing*, pp. 130–132, 2008.

[60] O. Amft, M. Stager, P. Lukowicz, and G. Tröster, "Analysis of chewing sounds for dietary monitoring," in *Ubiquitous Computing*, 2005.

[61] S. Paßler and W.-J. Fischer, "Food intake activity detection using a wearable microphone system," in *Intelligent Environments*, 2011.

[62] G. Shroff, A. Smailagic, and D. P. Siewiorek, "Wearable context-aware food recognition for calorie monitoring," in *IEEE International Symposium on Wearable Computers*, pp. 119–120, 2008.

[63] W. Fukuo, K. Yoshiuchi, K. Ohashi, H. Togashi, R. Sekine, H. Kikuchi, N. Sakamoto, S. Inada, F. Sato, T. Kadowaki, and A. Akabayashi, "Development of a hand-held personal digital assistant-based food diary with food photographs for japanese subjects," *Journal of the American Dietetic Association*, vol. 109, no. 7, pp. 1232–1236, 2009.

[64] A. A. Atienza, A. C. King, B. M. Oliveira, D. K. Ahn, and C. D. Gardner, "Using hand-held computer technologies to improve Dietary Intake," *American J. of Prev. Med.*, vol. 34, no. 6, pp. 514–518, 2008.

[65] D. B. Sharp and M. Allman-Farinelli, "Feasibility and validity of mobile phones to assess dietary intake," *Nutrition*, vol. 30, no. 11-12, pp. 1257–1266, 2014.

[66] L. Gemming, J. Utter, and C. Ni Mhurchu, "Image-assisted dietary assessment: a systematic review of the evidence.," *Journal of the Academy of Nutrition and Dietetics*, vol. 115, no. 1, pp. 64–77, 2015.

[67] A. D. Lassen, S. Poulsen, L. Ernst, K. K. Andersen, A. Biltoft-Jensen, and I. Tetens, "Evaluation of a digital method to assess evening meal intake in a free-living adult population," *Food and Nutrition Research*, vol. 54, no. 7, pp. 1–9, 2010.

[68] C. K. Martin, H. Han, S. M. Coulon, H. R. Allen, C. M, and S. D. Anton, "A novel method to remotely measure food intake of free-living people in real-time," *The British Journal of Nutrition*, vol. 101, no. 3, pp. 446–456, 2009.

[69] F. Zhu, M. Bosch, I. Woo, S. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp, "The use of mobile devices in aiding dietary assessment and evaluation.," *IEEE J. of Selected Topics in Sig. Proc.*, vol. 4, no. 4, pp. 756–766, 2010.

[70] O. Amft and G. Tröster, "Recognition of dietary activity events using on-body sensors," *Artificial Intelligence in Med.*, vol. 42, no. 2, pp. 121–36, 2008.

[71] H. Kalantarian, N. Alshurafa, and M. Sarrafzadeh, "A wearable nutrition monitoring system," in *Body Sensor Networks*, pp. 75–80, 2014.

[72] M. Farooq, P. C. Chandler-laney, M. Hernandez-reif, and E. Sazonov, "Monitoring of infant feeding behavior using a jaw motion sensor," *Journal of Healthcare Engineering*, vol. 6, no. 1, pp. 23–40, 2015.

[73] K. Corbin-Lewis and J. Liss, *Clinical anatomy & physiology of the swallow mechanism*. Cengage Learning, 2014.

[74] J. Lester, D. Tan, S. Patel, and A. Brush, "Automatic classification of daily fluid intake," *Pervasive Health*, pp. 1–8, 2010.

[75] A. Bedri, A. Verlekar, E. Thomaz, V. Avva, and T. Starner, "A Wearable System for Detecting Eating Activities with Proximity Sensors in the Outer Ear," in *Int. Symp. on Wearable Computers*, pp. 91–92, 2015.

[76] A. Bedri, A. Verlekar, E. Thomaz, V. Avva, and T. Starner, "Detecting mastication: A wearable approach," in *Int. Conf. on Multimodal Interaction*, pp. 247–250, ACM, 2015.

[77] T. Olubanjo, E. Moore, and M. Ghovanloo, "Unobtrusive and wearable systems for automatic dietary monitoring," *IEEE Transactions in Biomedical Engineering (Under Review)*, pp. 1–19, 2016.

[78] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, 2014.

[79] S. Youmans and J. Stierwalt, "An acoustic profile of normal swallowing," *Dysphagia*, vol. 20, no. 3, pp. 195–209, 2005.

[80] J. Cichero and B. Murdoch, "Acoustic signature of the normal swallow: characterization by age, gender and bolus volume," *Ann Otol Rhinol Laryngol*, vol. 111, no. 7, pp. 623–32, 2002.

[81] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin based feature selection - theory and algorithms," in *ICML*, 2004.

[82] H. C. Peng, F. H. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[83] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, no. 1-2, pp. 169–186, 2003.

[84] P. Lopez-Meyer, O. Makeyev, S. Schuckers, E. L. Melanson, M. R. Neuman, and E. Sazonov, "Detection of food intake from swallowing sequences by supervised and unsupervised methods.," *Annals of Biomedical Engineering*, vol. 38, no. 8, pp. 2766–74, 2010.

[85] P. Lopez-Meyer, S. Schuckers, O. Makeyev, J. Fontana, and E. Sazonov, "Automatic identification of the number of food items in a meal using clustering techniques based on the monitoring of swallowing and chewing," *BSPC*, vol. 7, no. 5, pp. 474–80, 2012.

[86] T. M. Nguyen, S. Ahuja, and Q. M. J. Wu, "A real-time ellipse detection based on edge grouping," in *IEEE SMC*, pp. 3280–6, 2009.

[87] M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A Food Recognition System for Diabetic Patients Based on an Optimized Bag-of-Features Model," *IEEE Journal of Biomed. and Health Informatics*, vol. 18, no. 4, pp. 1261–1271, 2014.

[88] H. Hoashi, T. Joutou, and K. Yanai, "Image recognition of 85 food categories by feature fusion," in *IEEE Int. Sym. of Multimedia*, 2010.

[89] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," in *IEEE Int. Conf. of Image Processing*, 2009.

[90] S. O'Hara and B. A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval," *ArXiv e-prints*, pp. 1–25, 2011.

[91] V. Bettadapura, E. Thomaz, A. Parnami, G. D. Abowd, and I. Essa, "Leveraging Context to Support Automated Food Recognition in Restaurants," *IEEE WACV*, pp. 580–587, 2015.

[92] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *IEEE Conf. on Computer Vision and Pattern Recognition - CVPR*, pp. 2249–2256, 2010.

[93] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features pre-trained with food-related categories," *IEEE Int. Conf. on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, 2015.

[94] W. Wu and J. Yang, "Fast food recognition from videos of eating for calorie estimation," in *IEEE Int. Conf. on Multimedia & Expo*, pp. 1210–3, 2009.

[95] A. Woda, A. Mishellany, and M. A. Peyron, "The regulation of masticatory function and food bolus formation," *Journal of Oral Rehabilitation*, vol. 33, no. 11, pp. 840–849, 2006.

[96] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[97] L. I. Smith, "A tutorial on Principal Components Analysis," *Cornell University, USA*, vol. 51, no. 52, p. 65, 2002.

[98] H. Boström, S. F. Andler, M. Brohede, R. Johansson, A. Karlsson, J. V. Laere, L. Niklasson, M. Nilsson, A. Persson, and T. Ziemke, "On the Definition of Information Fusion as a Field of Research," *IKI Technical Reports*, pp. 1–8, 2007.

[99] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[100] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "PFID: Pittsburgh fast-food image

dataset," in *IEEE ICIP*, 2009.

[101] B. Drake, "Food Crushing Sounds. An Introductory Study.," *J. of Food Science*, vol. 28, no. 2, 1963.

[102] M. Shiozawa, H. Taniguchi, H. Hayashi, K. Hori, T. Tsujimura, Y. Nakamura, K. Ito, and M. Inoue, "Differences in chewing behavior during mastication of foods with different textures," *Journal of Texture Studies*, vol. 44, no. 5, pp. 44–55, 2012.

[103] S. Matos, S. S. Birring, I. D. Pavord, and D. H. Evans, "Recordings using hidden markov models," *IEEE TBME*, vol. 53, no. 6, pp. 1078–1083, 2006.

[104] O. Amft and G. Tröster, "Methods for detection and classification of normal swallowing from muscle activation and sound," in *Pervasive Health*, pp. 1–10, 2006.

[105] S. Hamlet, R. Nelson, and R. Patterson, "Interpreting the sounds of swallowing: fluid flow through the cricopharyngeus," *Ann Otol Rhinol Laryngol*, vol. 99, pp. 749–52, 1990.

[106] J. A. Y. Cichero and B. E. Murdoch, "The physiologic cause of swallowing sounds: Answers from heart sounds and vocal tract acoustics," *Dysphagia*, vol. 13, no. 1, pp. 39–52, 1998.

[107] Z. Yan, V. Subbaraju, D. Chakraborty, A. Misra, and K. Aberer, "Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach," in *ISWC*, pp. 17 – 24, 2012.

[108] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *IEEE Computer Vision Workshops (ICCV)*, 2009.

[109] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 1194–1201, 2012.

[110] S. Stein and S. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Pervasive and Ubiquitous Computing*, pp. 729–738, 2013.

[111] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, and A. Batliner, "I hear you eat and speak: Automatic recognition of eating condition and food type, use-cases, and impact on ASR Performance," *PLoS ONE*, vol. 11, no. 5, pp. 1–24, 2016.

[112] D. Ramanan and D. a. Forsyth, "Automatic annotation of everyday movements," *Adv. in Neural Info. Proc. Sys. (NIPS)*, vol. 16, 2003.

[113] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "Annosearch: Image auto-annotation by search," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 1483–1490, 2006.

[114] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Computational Learning Theory*, pp. 92–100, 1998.

[115] D. Guan, W. Yuan, Y.-K. Lee, A. Gavrilov, and S. Lee, "Activity recognition based on semi-supervised learning," in *IEEE Embedded & Real-Time Computing Sys. and Apps.*, pp. 469–475, 2007.

[116] M. Stikic, D. Larlus, S. Ebert, and B. Schiele, "Weakly supervised recognition of daily life activities with wearable sensors," *Trans. on Pattern Analy. & Mach. Intell.*, vol. 33, no. 12, pp. 2521–37, 2011.

[117] G. Abowd, A. Dey, R. Orr, and J. Brotherton, "Context-awareness in wearable and ubiquitous computing," *Virtual Reality*, vol. 3, no. 3, pp. 200–211, 1998.

[118] D. Riboni and C. Bettini, "COSAR: Hybrid reasoning for context-Aware activity recognition," in *Personal and Ubiquitous Computing*, vol. 15, pp. 271–289, 2011.

[119] D. H. Hu and Q. Yang, "CIGAR: Concurrent & interleaving goal & activity recognition," in *AAAI Conf. on Artificial Intelligence*, pp. 1363–8, 2008.

[120] R. Helaoui, M. Niepert, and H. Stuckenschmidt, "Recognizing interleaved and concurrent activities: A statistical-relational approach," in *Pervasive Computing and Communications (PerCom)*, pp. 1–9, 2011.

[121] T. Winkler, "How realistic is artificially added noise?," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2605–2608, 2011.